university of groningen

Hidden Markov Models for the Analysis of Next-Generation-Sequencing Data

PhD Thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. E. Sterken
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

Monday 15 October 2018 at 16:15 hours

by

Aaron Sebastian Taudt

born on 12 January 1987
in Bergisch Gladbach, Germany

Supervisor
Prof. G. de Haan

Co-supervisor
Dr. M. Colomé-Tatché

Assessment committee
Prof. J.O. Korbel
Prof. B.J.L. Eggen
Prof. S. Ossowski

Paranymphs
David Roquis
Daniele Novarina

Preface

The creation of software that is useful for experimental biologists lies at the heart of this thesis. It was created in an environment where the majority of scientists conduct experimental research in the laboratory and only few perform purely bioinformatic research. This setting has allowed strong collaborations between method-developing bioinformaticians and wet-lab scientists, and has ensured that the methods presented in this thesis are not only *useful*, but are also *usable* by experimental scientists with little bioinformatic training.

In this thesis I present the algorithms and tools that I have developed in the course of my PhD. Many projects that I was involved in were highly collaborative efforts, and the experimental results which were obtained using the presented tools are presented in the form of their article abstracts at the end of each chapter in section “Applications”.

Chapter 1 gives a short introduction into Next Generation Sequencing (NGS) and shows how the chapters of this thesis are connected by NGS. It follows a didactic introduction to Hidden Markov Models, which is intended for readers who are yet unfamiliar with this technique. It will enable them to understand the concept of Hidden Markov Models as well as to easier understand the HMMs presented in later chapters of this thesis. It is intended to be easy to follow, but by no means general and exhaustive, and I refer the interested reader to [1] and [2] for a deeper understanding of HMMs.

Chapter 2 describes the `AneuFinder`, an algorithm for automated copy number detection from single-cell whole genome sequencing (scWGS) data. Although many tools for copy number detection already existed, none of them was specifically designed for single-cell sequencing data or provided the necessary quality control and heterogeneity analysis that people in ERIBA needed. This project is also a prime example of how the concept of “open space” in the ERIBA building has lead to very successful collaborations.

Chapter 3 presents a method for breakpoint detection and copy number analysis for (single-cell) Strand-seq data. The method is implemented in the `AneuFinder` package and makes use of the same pre-processing and post-processing options, but uses a different method for copy number calling and breakpoint detection. The methods described in this chapter can also be used for copy number calling in scWGS data (instead of the methods in Chapter 2), as they are superior in terms of robustness and accuracy. This reflects the two years of experience that lies between the development of both approaches.

Chapter 4 gives an introduction to chromatin states and describes an algorithm for the joint analysis of multiple ChIP-seq experiments. The algorithm is implemented in package `chromstaR`, and has become a versatile ChIP-seq analysis suite over the past 4 years. This project is also the source for many of the ideas that are used in other chapters of this thesis.

Chapter 5 describes `METHimpute`, a method for the analysis of whole genome bisulfite sequencing data. The algorithm was developed in the 4th year of my PhD, and it presents a very elegant implementation of a Hidden Markov Model.

Table of Contents

Chapter 1

Introduction

Aaron Taudt[1,2]

1. *European Research Institute for the Biology of Ageing, University of Groningen, University Medical Center Groningen, A. Deusinglaan 1, Groningen 9713 AV, The Netherlands.*
2. *Institute of Computational Biology, Helmholtz Zentrum München, Ingolstädter Landstr. 1, Neuherberg 85764, Germany.*

Next Generation Sequencing

The past decade has seen the rise of “Next Generation Sequencing (NGS)”, a powerful and versatile technology that allows the study of a wide range of nucleotide sequence-related phenomena. Traditional Sanger-sequencing was first applied in 1977 to sequence the ~5400 nucleotide long φX174 bacteriophage genome [3], and stayed the prevalent sequencing technology for the following 40 years. In the mid-2000s, the first NGS platforms were introduced and quickly superseded traditional Sanger-sequencing. Nowadays, large genomes like that of mammals or plants with billions of base pairs can be sequenced in a matter of days and a price approaching 1000$. The term “next generation” sequencing refers to the unprecedented scale at which sequencing is now possible. At the same time, Sanger sequencing continues to be the gold standard for sequencing and NGS validation because of its lower error rate.

A typical NGS workflow consists of DNA extraction and fragmentation, retention of desired fragments, size selection of fragments, adapter ligation (library preparation) and sequencing, alignment to the reference genome or de novo assembly of a genome. This modular workflow makes it possible to perform many different types of sequence-related experiments, depending on the type of fragments that are retained for sequencing. A schematic overview of the most common NGS workflows is depicted in Figure 1-1. Widely performed techniques are

- *whole genome sequencing*, where all DNA fragments are retained and sequenced;
- *ChIP-seq*[1], where only DNA interacting with a specific protein is retained via antibodies;
- *bisulfite-seq*, where DNA is treated with sodium bisulfite to convert unmethylated cytosines to uracil, which allows detection of methylated cytosines after sequencing;
- *exome sequencing*, where only protein coding regions are retained with specifically designed probes;
- *RNAseq*, where the total RNA content of a sample is subjected to sequencing;
- *Hi-C*, where DNA-DNA interactions are captured by cross-linking interacting DNA.

Each of these techniques gives a different view into the cell, and NGS has therefore been called a “molecular microscope” [4], emphasizing its importance and generality for biomedical research. NGS is now an essential tool for different fields such as genetics, epigenetics and clinical diagnostics [5]. At the same time, due to the similarity of the employed sequencing technology, the data output for each of the above techniques share similar features. This makes it possible to re-use bioinformatic solutions for one type of experiment in a modified form for the analysis of other types of experiments. The presented thesis is an example for this, in that it uses Hidden Markov Model-based approaches to analyze whole genome sequencing, ChIP-seq and bisulfite-seq data.

1 ChIP-seq: chromatin immunoprecipitation followed by sequencing

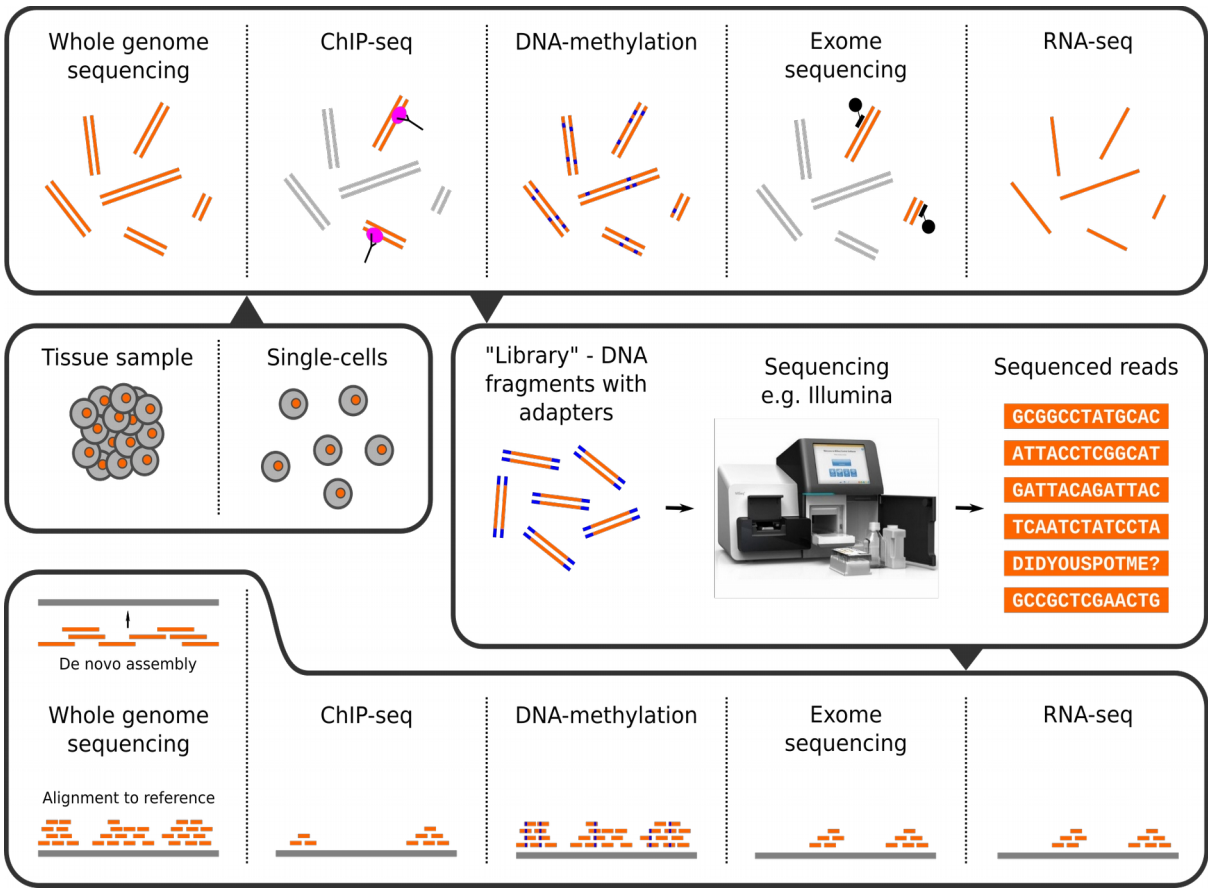


Figure 1-1 | Overview of the most common NGS techniques. Starting from a tissue sample or single-cells, DNA is extracted and purified. For *whole genome sequencing* (WGS), all extracted DNA is retained and sequenced. For *ChIP-seq*, only DNA that was cross-linked with a protein of interest is retained via specific antibodies. For assessment of *DNA-methylation* (Bisulfite sequencing), DNA is treated with sodium bisulfite to convert unmethylated cytosines to uracil, indicated as blue dots. For *exome sequencing*, only target DNA is retained with specifically designed probes and sequenced. For *RNA-seq*, RNA is extracted instead of DNA, reverse transcribed and sequenced. For all types of experiment, adapters must be added for the sequencing step. Finally, sequenced reads are either used for de novo assembly of a genome (only with WGS), or mapped to an existing reference genome.

A very recent development is the application of NGS to individual cells instead of bulk material. This requires modified protocols at the wet bench, but also new bioinformatic tools for analysis. The results from traditional bulk methods are an average over many cells, and therefore show continuous-valued behavior where single-cell methods show discrete-valued behavior. An example for this is the copy number of a chromosome, which can only have discrete values in a single cell, but which might be averaged to a real number depending on the level of heterogeneity between cells in a bulk experiment. The methods presented in Chapter 2 and 3 for copy number and breakpoint calling are explicitly developed for single cells. The methods in Chapter 4 and 5 for ChIP-seq peak calling and methylation status calling were developed for bulk experiments, but have a “single-cell structure” (by the use of a Hidden Markov Model with discrete states), potentially making them even more useful in the future when single-cell ChIP-seq and bisulfite-seq will be available.

Hidden Markov Models

Hidden Markov Models are a useful choice for data analysis whenever the following conditions are met:

- **The data is measured along one dimension.** This dimension can for instance be *time*, *spatial position* or *genomic coordinates*. Traditional applications of HMMs were mostly developed for time-series data, while bioinformatic HMMs are often developed with sequence position as measurement dimension.
- **Data points are correlated along the measurement dimension.** This is the case for many time-series signals, and also seems to be the case for most NGS data. By knowing one data point, one can thus make (probable) statements about the data points in the immediate vicinity.
- **There is a "hidden" truth in which one is interested, but which is not measured directly.** Instead, the measured signal is somehow related to the hidden truth. Examples for this hidden-vs-observed dichotomy might be the ploidy-state of a chromosome vs. the number of sequenced read fragments (Chapter 2), or the histone modification state vs. the height of the ChIP-seq signal (Chapter 4), or the methylation status of a cytosine vs. the bisulfite-seq signal (Chapter 5). To make things more concrete, let us look at a simple example.

An HMM for weather inference

To introduce the concept of Hidden Markov Models, it is helpful to first explain the concept of a Markov chain. Consider the women in Figure 1-2. She can observe the weather (for simplicity assumed to be either rainy or sunny) and note the weather for each day in a week. This chain of weather states is now called a Markov chain, if the weather on any given day solely depends on the weather of the previous day. This dependence on the previous day is modeled with a *transition probability*, which describes the probability of transitioning from one weather state to the next.

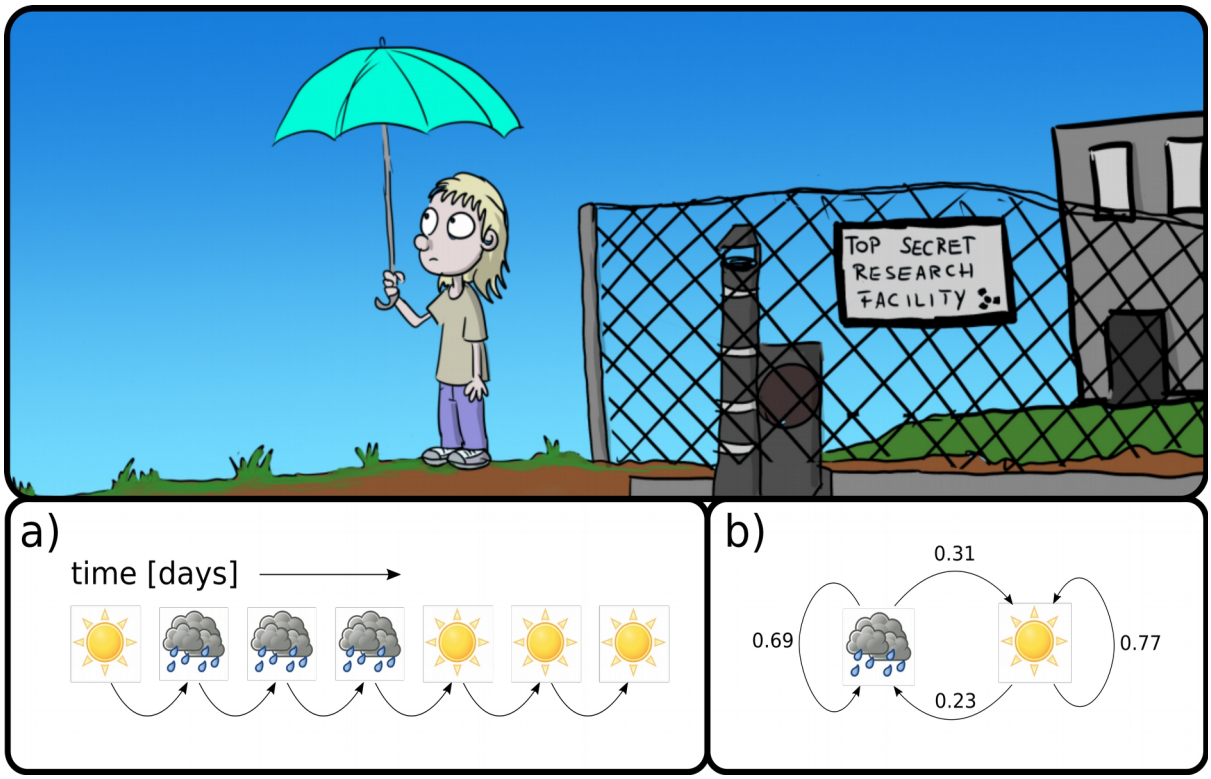


Figure 1-2 | Cartoon and schematic for a Markov chain. The woman with the umbrella can observe the weather and note down the "weather-state" for each day (**a**). For simplicity we assume that there are only two states – rainy and sunny. This weather chain is called a Markov chain if the weather on any given day solely depends (in a probabilistic way) on the weather of the previous day. **a** | The Markov chain and **b** | its graphical representation with transition probabilities. (Cartoon: Courtesy of monmon-comic.net by Simon Salomon)

Now that we have established the properties of a Markov chain, let us add one more layer of complexity and look at a Hidden Markov Model. The name "hidden" refers to the states of the model, which cannot be observed directly. Instead, the observation sequence is somehow correlated to the hidden states. The scientist in the underground facility in Figure 1-3 cannot observe the weather directly, *i.e.* the weather is hidden from him. However, knowing about the relationship between weather and air pressure, he can use a barometer to observe the average air pressure on each day and make inferences about the corresponding weather state outside. The relationship between air pressure and weather is depicted in the histogram in Figure 1-3|a. The distribution of air pressure values for rainy and sunny days is approximately normally distributed, and the distribution for sunny days has a higher mean than the distribution for rainy days. On sunny days, the air pressure is on average 10 hPa higher. However, although the means of the distributions differ, there is considerable overlap between air pressure on sunny and rainy days. Hence, just looking at the barometer will not tell the scientist with certainty about the outside weather. For example, observing a value of 1015 hPa will not tell the scientist whether it is rainy or sunny, since this value has quite a high probability on both rainy and sunny days. To improve the prediction, it is helpful to use the Markov chain property between daily weather-states, which means that the weather of one day depends on the weather of the previous day in a probabilistic way. An example is

given in Figure 1-3|b: Observing 1015 hPa on day 4 and day 7 will lead to different predictions, depending on the weather and air pressure the day before. This property of "borrowing" information from neighboring data points makes Hidden Markov Models an extremely useful tool for NGS sequencing analysis.

The structure of this Hidden Markov Model for weather inference is shown in Figure 1-3|c: It is fully defined by the number of states, their transition probabilities and their emission distributions. Here arises an important question: How do we know the structure of the HMM? The number of hidden states is usually dictated by the system under consideration, and reflects the information that one wishes to obtain from the system. In this case the underground scientist is interested in a classification of the weather into rainy and sunny days. Of course, he might also develop a more sophisticated model with three states – rainy, sunny, windy – or any other number of states that he is interested in. The shape of the emission distributions is usually determined with *a priori* knowledge about the nature of the observed variables (like in Figure 1-3|a), or an assumption about the relationship between hidden state and observed variable.

Now that we have defined the basic structure of an HMM, that is the number of hidden states and the shape of the emission distributions, we can continue to set the transition probabilities and emission distribution parameters. This can also be done with *a priori* knowledge, but far more often these parameters are estimated from the observed data in a process called *model training*. In the context of this work, we have used the Baum-Welch algorithm for model training[2]. In the Baum-Welch algorithm, model parameters are initialized with a random guess. Starting from this initial guess, the likelihood of observing the state-sequence with the current parameters is calculated, and parameters are then updated in order to maximize the likelihood function. The process of likelihood estimation and parameter updates is repeated until the likelihood stops changing.

To summarize, Hidden Markov Models are a useful method for classification tasks of one-dimensional data if there is correlation between neighboring data points and if the observed variable is correlated to an unobserved classification category (a hidden state). We have seen how a simple HMM can be used to infer the weather from air pressure values. An important difference of this example to many other introductory examples is the definition of the emission densities (air pressure distributions). In most HMM tutorials, hidden states emit *discrete symbols*, whereas in this example the hidden states emit values coming from a *continuous distribution*. This definition directly transfers to the models developed in later chapters of this thesis.

2 There are other methods for model training, *e.g.* Markov Chain Monte Carlo (MCMC). These methods are not used in the work presented in this thesis and are hence not covered in this introduction.

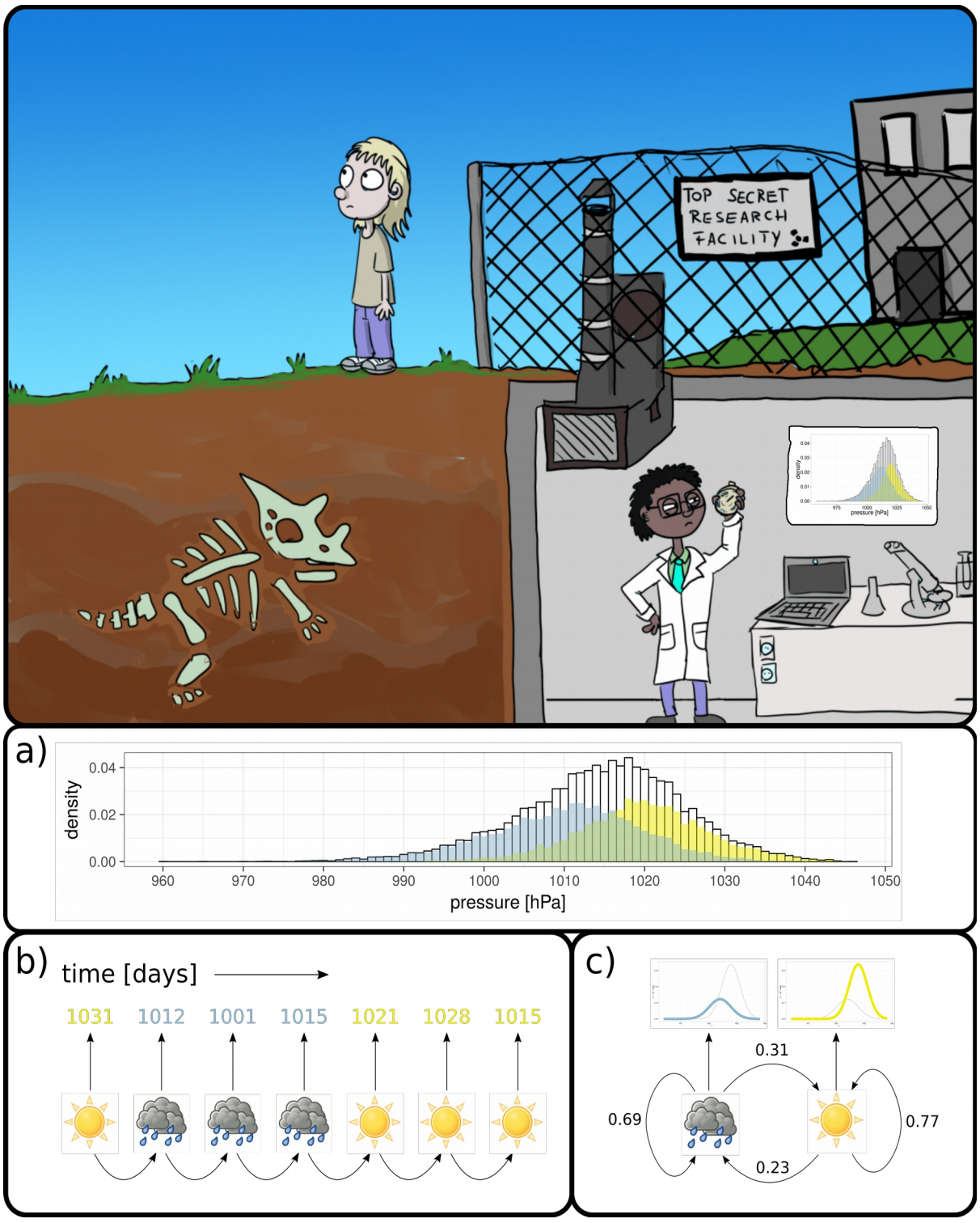


Figure 1-3 | Cartoon and schematic for a Hidden Markov chain. The scientist in the underground research facility cannot observe the weather directly. However, since he knows about the connection between air pressure and weather (**a**), he can use a barometer to make inferences about the outside weather. **a** | Histogram of average daily air pressure values, colored by rainy and sunny days. The data is from a weather station in Leeuwarden from 1951 to 2015 and days without precipitation are defined as sunny (source: https://www.knmi.nl/nederland-nu/klimatologie/daggegevens). **b** | The Hidden Markov chain with emitted air pressure values. **c** | The graphical representation of the HMM with transition probabilities and emission densities. (Cartoon: Courtesy of monmon-comic.net by Simon Salomon)

Mathematical notation

We will follow notation for variable names introduced in [6] in this thesis. Additionally, scalar variables X are written in italic, while anything higher-dimensional such as vectors or matrices $\boldsymbol{X}$ are written in bold. Vector or matrix elements X_{ij} or $X_{i,j}$ will be selected with subscript indices, and these indices will usually be one letter indices, possibly separated by commas for readability.

A Hidden Markov Model is completely characterized by three elements: 1) the initial probabilities for being in any one state $\boldsymbol{\pi}$; 2) the transition matrix $\boldsymbol{A}$, which contains the possible transitions from any state i to any other state j; 3) and the emission distributions $\boldsymbol{B}$ for each state i. We might summarize this set of parameters with $\boldsymbol{\lambda} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$. We assume that our HMM has N states and our observed sequence $\boldsymbol{y}$ has T elements. We will usually iterate/sum over the states with i or j and over the sequence with t.

We begin by writing down the likelihood P for the observed sequence $\boldsymbol{y}$. This is obviously the product of the likelihoods of being in state i at each observation t, and emitting observation value B_{it}:

$$P(\boldsymbol{y}) = \underbrace{\prod_{i}^{N} \pi_i^{z_{i,1}}}_{\text{initial state}} \underbrace{\prod_{i}^{N}\prod_{j}^{N}\prod_{t}^{T-1} A_{ij}^{z_{i,t}z_{j,t+1}}}_{\text{all other states}} \underbrace{\prod_{i}^{N}\prod_{t}^{T} B_{it}^{z_{i,t}}}_{\text{observed value}} \qquad \text{(eq. 1.1)}$$

We made use of a "component indicator" $\boldsymbol{z}_t$ to make sure that only terms for the correct hidden state are used when multiplying the probabilities. This component indicator vector is 1 for the true hidden state and 0 for all other states. Figure 1-4|b shows how this component indicator looks like for our weather HMM.

Taking the logarithm of this expression we obtain the log-likelihood L:

$$L = \underbrace{\sum_{i}^{N} z_{i,1}\log(\pi_i)}_{\text{initial state}} + \underbrace{\sum_{i}^{N}\sum_{j}^{N}\sum_{t}^{T-1} z_{i,t}z_{j,t+1}\log(A_{ij})}_{\text{all other states}} + \underbrace{\sum_{i}^{N}\sum_{t}^{T} z_{i,t}\log(B_{it})}_{\text{observed value}} \quad \text{.(eq. 1.2)}$$

Since in reality the hidden state is usually unknown, and thus we cannot know the component indicator vectors $\boldsymbol{z}_t$, we help ourselves by replacing the component indicator vectors with their conditional probabilities:

$$z_{i,t} \rightarrow P(z_{i,t}=1 \mid \boldsymbol{y}) = \gamma_{i,t} \quad \text{and} \qquad \text{(eq. 1.3)}$$

$$z_{i,t}z_{j,t+1} \rightarrow P(z_{i,t}=1, z_{j,t+1}=1 \mid \boldsymbol{y}) = \xi_{i,j,t} \quad . \qquad \text{(eq. 1.4)}$$

The variable γ_{it} is the probability of being in state i at position t, given the observed sequence $\boldsymbol{y}$. This variable is also called *posterior probability* and is used for inference of the hidden state, for example by maximizing over the different values $\boldsymbol{\gamma}_i$ for each t. ξ_{ijt} is the probability of being in state j at t+1, and having been in state i at t. In order to calculate $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$, it is useful to define two more variables $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, called forward and backward variables.

The forward variable $\boldsymbol{\alpha}_t$ gives the probability of the partial observation sequence from the beginning until t, and the backward variable $\boldsymbol{\beta}_t$ gives the probability of the partial observation sequence from $t+1$ until the end. A visual aid for understanding $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is given in Figure 1-4|c. Forward and backward variables are calculated inductively and calculation is described in section “Implementing an HMM” (page 14).

By replacing the component indicators with their conditional probabilities, the log-likelihood becomes the conditional expectation Q. This is the governing equation from which one can derive all updating formulas for the Baum-Welch algorithm:

$$Q = \underbrace{\sum_{i}^{N} \gamma_{i,t=0} \log(\pi_i)}_{\text{initial probabilities}} + \underbrace{\sum_{i,j,t}^{N,N,T-1} \xi_{ijt} \log(A_{ij})}_{\text{transition probabilities}} + \underbrace{\sum_{i,t}^{N,T} \gamma_{it} \log(B_{it})}_{\text{emission distributions}} \qquad \text{(eq. 1.5)}$$

Updates for any HMM parameter x are now obtained by “simply” solving the partial derivative of Q for x:

$$\frac{\partial Q}{\partial x} = 0 \quad . \qquad \text{(eq. 1.6)}$$

The solution of this step depends very much on the specific implementation of the HMM, and is hence described in each chapter separately.

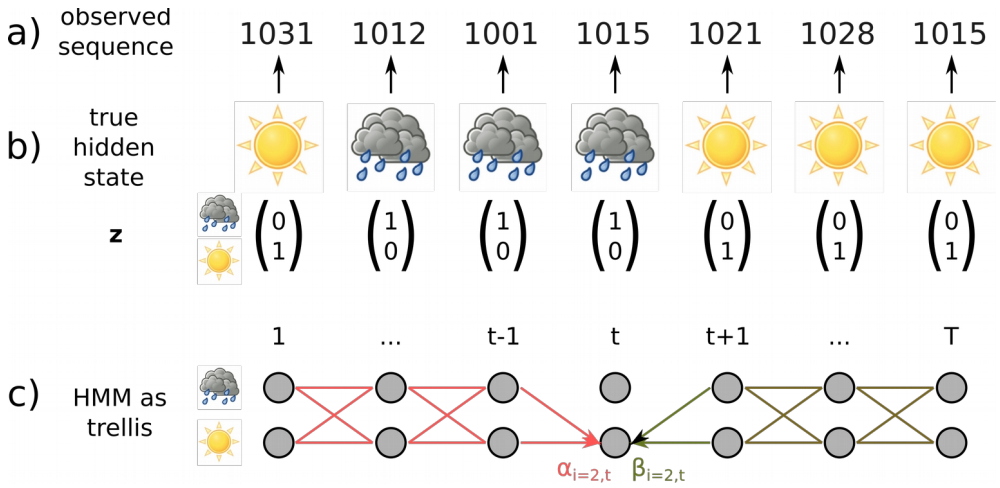


Figure 1-4 | Visual aid for HMM variables. **a** | Observed sequence $\boldsymbol{y}$ and **b** | hidden truth with component indicator vectors $\mathbf{z}_t$. **c** | Trellis for the HMM with colored lines visualizing the concept of forward α and backward β variables.

Implementing an HMM

This section summarizes the formulas which are needed to actually implement a Hidden Markov Model. It will be useful for everyone who is planning to implement their own HMM and might be safely skipped by everyone else.

Forward variables $\boldsymbol{\alpha}_t$ give the probability of the partial observation sequence from beginning until t (see Figure 1-4|c), and can be obtained inductively with:

1. Initialization: $\alpha_{i,t=1} = \pi_i B_{i,t=1} \quad , \; 1 \le i \le N$ (eq. 1.7)

2. Induction: $\alpha_{i,t} = \left[\sum_{j=1}^{N} \alpha_{j,t-1} A_{j,i} \right] B_{i,t} \quad , \; 2 \le t \le T, \; 1 \le i \le N$ (eq. 1.8)

Backward variables $\boldsymbol{\beta}_t$ give the probability of the partial observation sequence from t+1 until the end given state i at time t (see Figure 1-4|c), and can be obtained inductively with:

1. Initialization: $\beta_{i,t=T} = 1 \quad , \; 1 \le i \le N$ (eq. 1.9)

2. Induction: $\beta_{i,t} = \sum_{j=1}^{N} A_{i,j} B_{j,t+1} \beta_{j,t+1} \quad , \; 1 \le t \le T-1, \; 1 \le j \le N$ (eq. 1.10)

The likelihood $P(\boldsymbol{y} \mid \boldsymbol{\lambda})$ of observing a given sequence $\boldsymbol{y}$ can be computed as either of

$$P(\boldsymbol{y} \mid \boldsymbol{\lambda}=(\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})) = \sum_{i=1}^{N} \alpha_{i,T} = \sum_{i=1}^{N} \pi_i B_{i,t=1} \beta_{i,t=1} = \sum_{i=1}^{N} \alpha_{i,t} \beta_{i,t} \qquad \text{(eq. 1.11)}$$

Two other useful variables are the posterior probability $\boldsymbol{\gamma}_t$ and another variable $\boldsymbol{\xi}_t$. The variable γ_{it} is the probability of being in state i at position t, given the observed sequence $\boldsymbol{y}$. The ξ_{ijt} is the probability of being in state j at t+1, and having been in state i at t. They can be calculated as

$$\gamma_{i,t} = P(i \mid \boldsymbol{\lambda}) = \frac{\alpha_{i,t} \beta_{i,t}}{P(\boldsymbol{y} \mid \boldsymbol{\lambda})} \quad \text{and} \qquad \text{(eq. 1.12)}$$

$$\xi_{i,j,t} = P(i \text{ at } t, j \text{ at } t+1 \mid \boldsymbol{\lambda}) = \frac{\alpha_{i,t} A_{i,j} B_{j,t+1} \beta_{i,t+1}}{P(\boldsymbol{y} \mid \boldsymbol{\lambda})} \quad . \qquad \text{(eq. 1.13)}$$

They satisfy the relationship $\gamma_{i,t} = \sum_{j=1}^{N} \xi_{i,j,t}$. (eq. 1.14)

The above formulas are given for understanding the definition of these variables. However, any practical HMM implementation must scale these variables appropriately. The reason for this is that we are multiplying probabilities which are between 0 and 1 and the forward and backward variables will thus quickly approach zero. The precision required for these calculations exceeds the capability of computer processors. Therefore, it is necessary to do proper scaling at each step. A commonly used scaling scheme for the forward variables is:

1. Initialization:

$$\ddot{\alpha}_{i,t=1} = \alpha_{i,t=1} \quad \text{(eq. 1.15)}$$

$$c_{t=1} = 1 / \sum_{i=1}^{N} \ddot{\alpha}_{i,t=1} \quad \text{(eq. 1.16)}$$

$$\hat{\alpha}_{i,t=1} = c_{t=1}\ddot{\alpha}_{i,t=1} \quad \text{(eq. 1.17)}$$

2. Induction:

$$\ddot{\alpha}_{i,t} = \left[\sum_{j=1}^{N} \hat{\alpha}_{j,t-1} A_{j,i}\right] B_{i,t} \quad , \; 2 \le t \le T, \; 1 \le i \le N \quad \text{(eq. 1.18)}$$

$$c_t = 1 / \sum_{i=1}^{N} \ddot{\alpha}_{i,t} \quad \text{(eq. 1.19)}$$

$$\hat{\alpha}_{i,t} = c_t \ddot{\alpha}_{i,t} \quad \text{(eq. 1.20)}$$

The coefficient c_t is the scaling factor and there is one scaling factor for each step t. This same scaling factor can be used to scale the backward variables and does not need to be re-computed. The scaling scheme for the backward variables looks now like this:

1. Initialization:

$$\ddot{\beta}_{i,t=T} = 1 \quad \text{(eq. 1.21)}$$

$$\hat{\beta}_{i,t=T} = c_{t=T}\ddot{\beta}_{i,t=T} \quad \text{(eq. 1.22)}$$

2. Induction:

$$\ddot{\beta}_{i,t} = \sum_{j=1}^{N} A_{i,j} B_{j,t+1} \hat{\beta}_{j,t+1} \quad , \; 1 \le t \le T-1, \; 1 \le j \le N \quad \text{(eq. 1.23)}$$

$$\hat{\beta}_{i,t} = c_t \ddot{\beta}_{i,t} \quad \text{(eq. 1.24)}$$

This scaling approach fulfills some useful relationships, which can be used to rewrite the γ_{it} and ξ_{it} and the derived updating formulas for the Baum-Welch algorithm with the scaled variables:

$$\sum_{i=1}^{N} \hat{\alpha}_{i,t} = 1 \quad \text{(eq. 1.25)}$$

$$\alpha_{i,t} = \hat{\alpha}_{i,t} / \left(\prod_{\tau=1}^{t} c_\tau\right) = \hat{\alpha}_{i,t} / C_t \;\; \text{with} \;\; \prod_{\tau=1}^{t} c_\tau = C_t \quad \text{(eq. 1.26)}$$

$$\beta_{i,t} = \hat{\beta}_{i,t} / \left(\prod_{\tau=t}^{T} c_\tau \right) = \hat{\beta}_{i,t} / D_t \quad \text{with} \quad \prod_{\tau=t}^{T} c_\tau = D_t \qquad \text{(eq. 1.27)}$$

$$P(\boldsymbol{y} \mid \boldsymbol{\lambda}) = 1 / \prod_{t=1}^{T} c_t \quad \text{and} \quad \log P(\boldsymbol{y} \mid \boldsymbol{\lambda}) = -\sum_{t=1}^{T} \log c_t \qquad \text{(eq. 1.28)}$$

$$C_t D_{t+1} = C_T = \frac{1}{P(\boldsymbol{y} \mid \boldsymbol{\lambda})} \quad \text{and} \quad C_t D_t = C_T c_t = \frac{c_t}{P(\boldsymbol{y} \mid \boldsymbol{\lambda})} \qquad \text{(eq. 1.29)}$$

Chapter 2

`AneuFinder`: An HMM for copy number detection in single-cell whole genome sequencing

Aaron Taudt[1,2], Bjorn Bakker[1], David Porubsky[1], Hilda van den Bos[1], Diana C. J. Spierings[1], Victor Guryev[1], Peter M. Lansdorp[1,3], Floris Foijer[1], Maria Colomé-Tatché[1,2]

1. *European Research Institute for the Biology of Ageing, University of Groningen, University Medical Center Groningen, A. Deusinglaan 1, Groningen 9713 AV, The Netherlands.*
2. *Institute of Computational Biology, Helmholtz Zentrum München, Ingolstädter Landstr. 1, Neuherberg 85764, Germany.*
3. *Terry Fox Laboratory, BC Cancer Agency, Vancouver, BC V5Z 1L3, Canada. Division of Hematology, Department of Medicine, University of British Columbia, Vancouver, BC V6T 1Z4, Canada.*

Based on Genome Biology 2016; doi: 10.1186/s13059-016-0971-7

Abstract

Aneuploidy is an aberrant number of chromosomes in a cell. This is associated with cognitive and developmental defects, and has been implicated to play a role in Alzheimer's disease. Aneuploidy is also a hallmark of cancer cells, and roughly two out of three tumors show aneuploid karyotypes. Furthermore, cancer cells show also smaller copy number alterations (CNA) – duplications or deletions of parts of a chromosome – and these are thought to contribute to tumor evolution and treatment success. Single-cell whole genome sequencing (scWGS) is a novel technique that allows to map these copy number alterations in non-dividing single cells. This chapter presents a computational method for mapping and analyzing CNA and aneuploidies from scWGS data, based on a Hidden Markov Model (HMM). The HMM models every copy number as a distinct hidden state and uses negative binomials as emission distributions. Measures for assessing library quality and karyotype heterogeneity are also presented.

Introduction

A typical mammalian cell has two complete sets of chromosomes, one inherited from the father and one from the mother. This state is called diploid and it implies that the whole genome is present in duplicate. Therefore, also every gene comes with two copies, and these gene-copies are termed alleles. Alleles are not necessarily identical, since one gene-copy was inherited from the mother and the other gene-copy was inherited from the father. During cell division (mitosis), chromosomes are replicated and distributed to the daughter cells, and if everything works well, each daughter cell will end up with a complete set of diploid chromosomes. The spindle assembly checkpoint (SAC) plays an important role in this process and ensures that chromosomes are distributed correctly into the daughter cells. It does this by blocking mitotic cells in metaphase until proper kinetochore-microtubule attachment and tension have been established. The SAC can be experimentally disturbed, for instance by truncating the SAC kinase Mps1, and this will lead to chromosome mis-segregation and cells with an abnormal number of chromosomes, a state called aneuploid. The condition in which cells continuously mis-segregate chromosomes during cell division is termed chromosomal instability (CIN). When whole chromosomes have aberrant copy numbers, we speak of aneuploidy, but CIN can also provoke structural copy number aberrations (CNA), where only parts of the chromosome or only some genes have an unphysiological number of copies.

CIN and the resulting aneuploidy have been shown to cause physiological stress and growth defects in yeast and primary mouse embryonic fibroblasts. Furthermore, some of the mouse models that were engineered to model CIN are characterised with a reduced lifespan, which can be rescued by reducing the levels of CIN. Although aneuploidy has detrimental consequences for untransformed cells, more than two out of three cancers are aneuploid, suggesting a fundamental relationship between aneuploidy and tumorigenesis that so far remains poorly understood. Aneuploidy is also associated with cognitive and developmental defects, the most prominent example being Down’s syndrome, which is caused by an additional copy of chromosome 21.

Single-cell whole genome sequencing (scWGS) is a novel technique to assess aneuploidy and copy number aberrations in single cells. Traditional metaphase spread-based techniques like FISH or Giemsa staining are limited to dividing cell populations, while other techniques like interphase FISH are limited in their resolution (only a few chromosomes can be probed), and yet other techniques like array CGH are only practicable in bulk experiments [7]. Single-cell whole genome sequencing combines the best of all worlds and is able to probe copy numbers at high resolution in non-dividing single cells. In a typical library preparation protocol, DNA is fragmented, single-stranded overhangs are repaired to blunt ends, then A-tailed and ligated with adapters necessary for the sequencing procedure. To increase the amount of DNA, a PCR amplification step might be added after adapter ligation, but this also introduces

additional biases in the sequencing. Multiple samples can be analyzed in one sequencing lane by adding barcodes in the adapter sequences, a process called multiplexing. After sequencing, libraries are demultiplexed, quality controlled, and reads are mapped to a reference genome.

In order to automatically detect aneuploidies and copy number aberrations from the mapped reads, as well as to perform reliable quality control of the obtained sequencing results, we have developed `AneuFinder`, an automated analysis pipeline with the following key features: 1) Independence of an external reference for copy number analysis; 2) Automated quantification of CNAs using a Hidden Markov Model; 3) Stringent semi-automated quality control of individual sequencing libraries; 4) Definition of measures for the assessment of karyotype heterogeneity and aneuploidy.

Model specification

Copy number detection with `AneuFinder` consists of three steps: (1) Binning (Figure 2-1|b), (2) correction for GC content (Figure 2-2), and (3) copy number detection with a Hidden Markov Model (Figure 2-1|c-d). This is followed by a semi-automated quality control for all libraries using a multivariate clustering approach (Figure 2-1|e). Finally, heterogeneity and aneuploidy scores are calculated (Figure 2-1|f).

Binning

We implemented two different binning strategies, fixed-width and variable-width windows. In the fixed-width binning strategy, we partition the genome into T non-overlapping, equally sized bins (default 1Mb) and count the number of aligned reads that overlap any given bin t. The variable-width binning requires a euploid reference that can either be a simulated or a real reference (*e.g.* many merged euploid single cell libraries). The bins are constructed as follows: 1) The euploid reference is binned into fixed-width windows of a given size (default 1Mb) and reads are counted in each bin. 2) The mode of read counts per fixed-width bin (X) is taken as the desired number of reads for the variable-width bins. 3) Variable-width bins are constructed such that each bin contains X reference reads (*i.e.* reads in the euploid reference). Fixed-width windows can lead to artifacts in the form of low copy number states that are caused by low-mappability regions. We therefore conducted all analyses in [8] with the variable-width binning approach, which partly corrects for mappability bias. Reference files for [8] were generated by merging reads from 46 diploid single cells for mouse (thymus T320) and 52 diploid single cells for human [9].

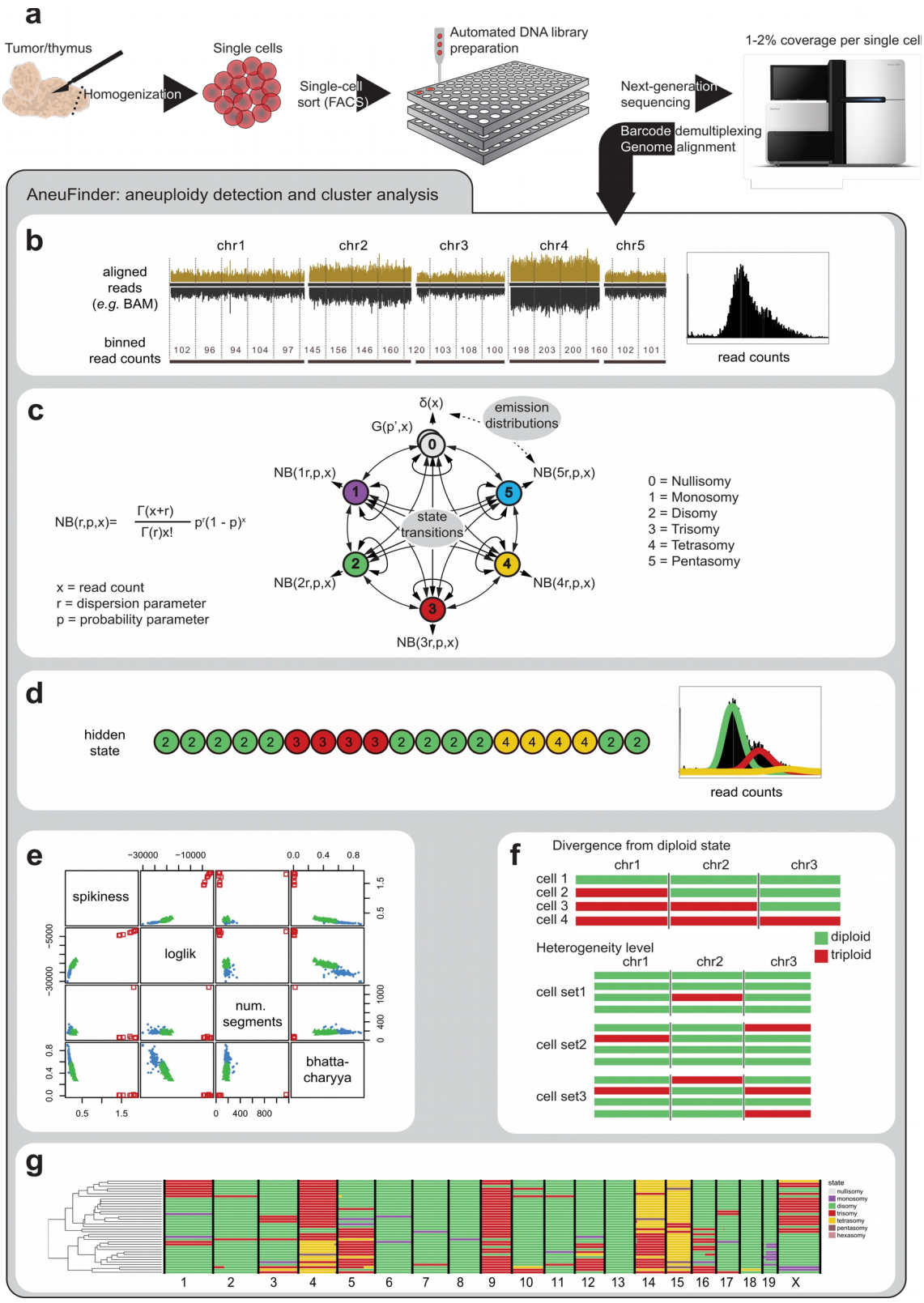


Figure 2-1 | AneuFinder – automated copy number analysis of single-cell sequencing data. **a** | Samples are homogenised, single-cell sorted and sequenced. **b** | Aligned sequencing reads are counted in non-overlapping bins of variable size based on mappability. **c** | A Hidden Markov Model with multiple hidden states is applied to the binned read counts in order to predict copy number state of every single bin. Emission distributions are modeled as negative binomial distributions (NB(*r*, *p*, *x*)). **d** | The model parameters are estimated using the Baum Welch algorithm and every binned read count is assigned to the copy number state that maximises the posterior probability. **e** | Quality of each single-cell library is assessed based on the following measures: spikiness, loglikelihood of the model determined by the Baum-Welch algorithm, number of separate copy

number segments and Bhattacharyya distance. Libraries are clustered based on these measures: the highest scoring cluster is selected for further analysis. **f** | The extent of aneuploidy is measured as the divergence of a given chromosome from the normal euploid state. At the cell population level, heterogeneity is measured as the number of cells with a distinct copy number profile within the population. **g** | example of a genome-wide copy number profile of a population of T-ALL cells. Each row represents a single cell with chromosomes plotted as columns. Copy number states are depicted in different colours. Cells are clustered based on the similarity of their copy number profile. (Source: Bakker and Taudt et al. 2016, [8])

Blacklist

Variable-width bins offer a partial correction for mappability, however, even with variable-width bins we could still observe artifacts around centromeric regions, caused by an extremely high number of mis-mapped reads to these regions. We chose a blacklisting strategy to exclude reads from artifact-prone regions from the analysis. Blacklists for [8] were generated by binning the references into fixed-width bins of 100 kb and blacklisting all bins where the read count was above the 0.9985 quantile or below the 0.1 quantile, respectively.

GC correction[3]

Binned read count values have been observed to have a unimodal relationship with GC content, where regions with high or low GC content have decreased read count values, compared to regions with intermediate GC content [10]. This makes it important to correct the read count values for GC bias, because otherwise the predicted copy number states might be confounded with GC content and would not necessarily reflect the correct copy number state. In order to achieve that, we partition the genome into T non-overlapping, variable or fixed-width bins as described above and count the number of aligned reads that fall into any given bin t [11]. This read count x_t is further GC-corrected with a model modified from [10]. For every bin t the GC content is determined as a fraction between 0 and 1 by counting the number of C and G in the bin and dividing by the bin size. The read count x_t is then multiplied by a correction factor f_{GC} that is dependent on the GC content:

$$x_{t,corrected} = x_t \cdot f_{GC} \qquad \text{(eq. 2.1)}$$

To calculate the correction factor, we group bins with GC content in one of 20 equally spaced intervals between 0 to 1 and calculate a correction factor f'_{GC} as follows

$$f'_{GC} = \frac{mean(x_{global})}{mean(x_{GC})} \qquad \text{(eq. 2.2)}$$

3 Please note that this section describes the method for GC-correction that was used in Genome Biology 2016; 10.1186/s13059-016-0971-7 and has meanwhile been replaced by a more powerful method described in Chapter 3.

where mean(x_{global}) is the average read count over all bins and mean(x_{GC}) is the average read count for all windows with a GC content in one of 20 equally spaced intervals between 0 to 1 (we use a trimmed mean, omitting 5% of bins from both extremes). Finally, we obtain a smoothed correction factor f_{GC} by fitting a second order polynomial to the correction factor f'_{GC}. This smoothing is done to obtain a good correction factor for bins with a very rare GC content. A second order polynomial was chosen because of the unimodal relationship between GC content and count values. Corrected read counts are rounded to the nearest integer.

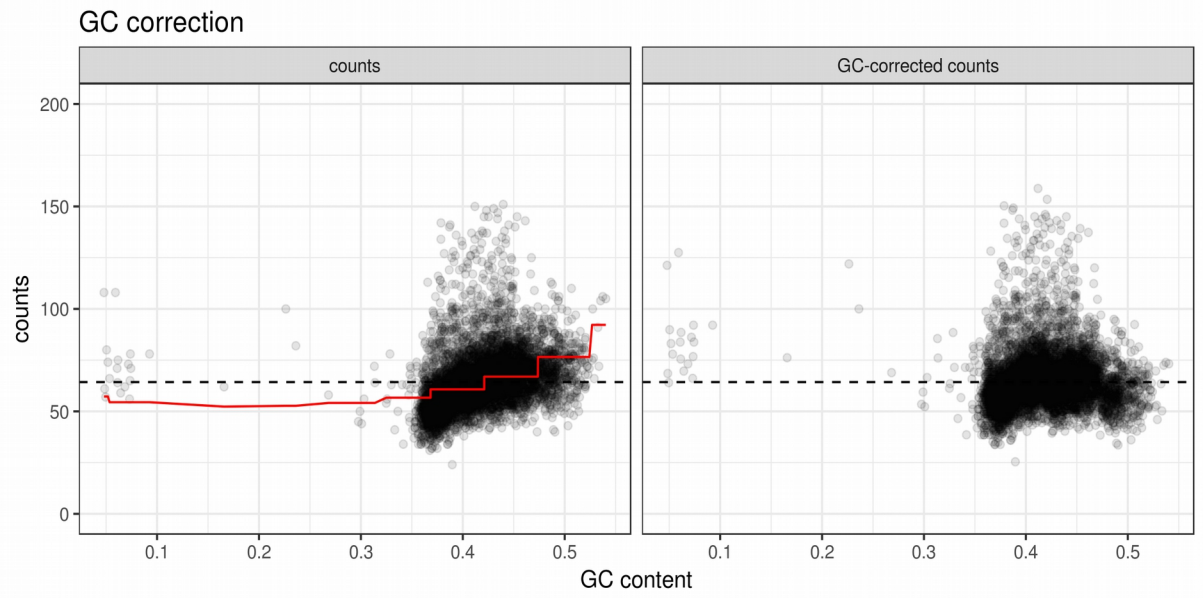


Figure 2-2 | GC correction for a typical single-cell. Raw read counts (left panel) show a GC-bias, which can be partly corrected for by the proposed correction procedure (right panel). The black dashed line indicates the mean read count mean(x_{global}), and the red line visualizes the correction factor f_{GC} (displayed is mean(x_{global}) / f_{GC}). (Cell ID = MM150218_I_22.bam, binsize = 400 kb).

Hidden Markov Model

A Hidden Markov Model is a good choice to infer (hidden) copy numbers from (observed) read count values, because we can assume that neighboring bins with the same copy number have correlated read count values, and that each copy number state has a unique distribution of associated read count values. Particularly, we can assume that the count values for a region with two copies are on average twice as high as the count values for a region with only one copy. Furthermore, we also assume that the variance of the observed read counts is twice as high for a region with two copies as compared to a region with one copy. We found negative binomial distributions to be a good empirical fit for the observed read count values (see Figure 2-3|b).

In our HMM, each copy number is modeled by a distinct hidden state. The read count distribution for each copy number ≥ 1 is modeled by a negative binomial distribution:

$$NB(r, p, x_t) = \frac{\Gamma(x_t + r)}{\Gamma(r)\, x_t!}\, p^r (1-p)^{x_t} \quad , \qquad \text{(eq. 2.3)}$$

where Γ is the Gamma function. The probability parameter p is assumed to be the same for all states ≥ 1 (monosomy, disomy, trisomy, etc.) and the dispersion parameters r are assumed to be multiples of the dispersion parameters for state 1 (monosomy). These choices for r and p are identical with setting mean and variance for each copy number as multiples of the mean and variance of state 1 (monosomy). Please see also Figure 2-1|c for a graphical representation of the HMM.

Nullisomies (copy number 0) are modeled by two hidden states: One "zero-inflation" state with a delta distribution to model gaps where no reads can be aligned and one state with a geometric distribution to account for mis-mapping reads in regions with zero copies:

$$Geom(p, x_t) = p(1-p)^{x_t} \qquad \text{(eq. 2.4)}$$

Model parameters are fitted with the Baum-Welch algorithm [1]. Compared to a standard HMM, the interconnected distribution parameters require modified updating formulas. The derivation of the modified updating formulas is detailed below, and uses notation introduced in [6]. Please see section "Mathematical notation" in the introduction for details about the notation.

The conditional expectation Q that needs to be maximized can be written as

$$Q = \sum_{i}^{N} \gamma_{i,t=0} \log(\pi_i) + \sum_{i,j,t}^{N,N,T-1} \xi_{ijt} \log(A_{ij,c_{t,t+1}}) + \sum_{i,t}^{N,T} \gamma_{it} \log(B_{it}) \quad . \qquad \text{(eq. 2.5)}$$

The updated parameters for the negative binomial distributions can be obtained by solving $\frac{\partial Q}{\partial p} = 0$ and $\frac{\partial Q}{\partial r} = 0$.

For independent negative binomial distributions, this would yield

$$p_i = \left(\sum_{t}^{T} \gamma_{it} \cdot r_i\right) / \left(\sum_{t}^{T} \gamma_{it} \cdot (r_i + x_t)\right) \quad \text{and} \qquad \text{(eq. 2.6)}$$

$$\frac{\partial Q}{\partial r_i} = \sum_{t}^{T} \gamma_{it} \cdot \left(\log(p_i) - \Psi(r_i) + \Psi(r_i + x_t)\right) = 0 \quad , \qquad \text{(eq. 2.7)}$$

where $\Psi(x) = \frac{\partial}{\partial x} \log \Gamma(x)$ is the digamma function. The equation for r_i cannot be solved analytically, but can be solved with a numerical Newton-Raphson method to obtain the updated parameters.

For the case $p_i = p$ where the probability parameter is the same for all states, and the dispersion parameters $r_i = i \times r$ are multiples of the dispersion parameter for the monosomy-state, we obtain instead

$$p = \left(\sum_{i,t}^{N,T} \gamma_{it} \cdot r\right) / \left(\sum_{i,t}^{N,T} \gamma_{it} \cdot (r + x_t)\right) \quad \text{and} \qquad \text{(eq. 2.8)}$$

$$\frac{\partial Q}{\partial r} = \sum_{i,t}^{N,T} \gamma_{it} \cdot \left(\log(p) - \Psi(i \cdot r) + \Psi(i \cdot r + x_t)\right) = 0 \quad . \qquad \text{(eq. 2.9)}$$

Again, a numerical Newton-Raphson method was implemented to solve equation 2.9.

The update for the geometric distribution is simply $p = \left(\sum_{t}^{T} \gamma_{it}\right) / \left(\sum_{t}^{T} \gamma_{it} \cdot (1 + x_t)\right)$.

Finally, after convergence of the Baum-Welch algorithm, the copy number state i_t is determined by maximizing over the posterior probabilities $i_t = \text{argmax}_i(\gamma_{it})$.

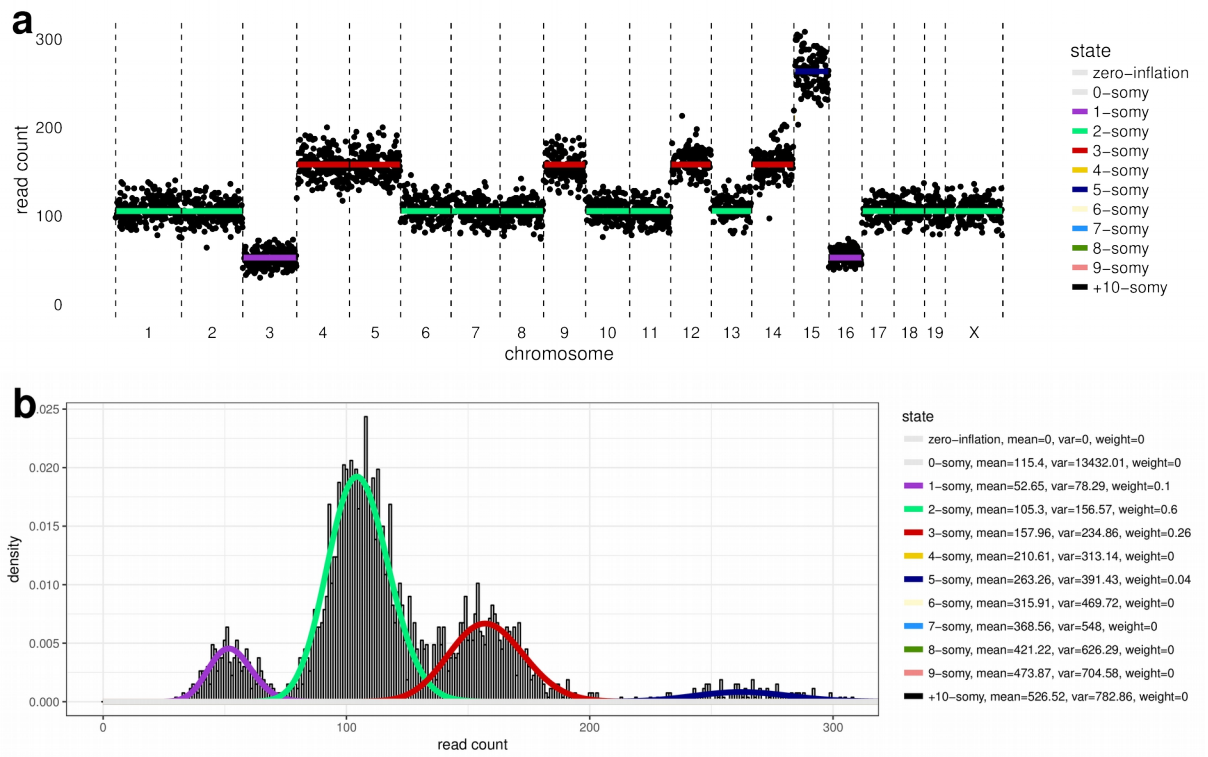


Figure 2-3 | AneuFinder plots for a typical single cell. The cell has a coverage of 0.6% (0.32 million reads, mouse genome) and was analyzed at bin size 1 Mb. **a** | Copy number profile with chromosomes on the x-axis and read count on the y-axis. Each dot represents a bin. **b** | Histogram of read count values and fitted distributions. At this bin size, distributions of the different copy number states are nicely separable. (Library ID = BB140512_III_006.bam, binsize = 1 Mb).

Quality control

Quality control has to be an integral part of the analysis pipeline, since single-cell sequencing libraries can be inherently noisy. The first approach one would think off to discern good-quality from bad-quality libraries might be a simple cutoff on the total number of sequenced reads, because libraries with more reads and hence deeper coverage would contain more information than libraries with shallow coverage. However, such a simple cutoff on the number of sequenced reads turns out to be too simplistic. This is evidenced in Figure 2-4|a, where two libraries with an identical number of sequenced reads are shown which are obviously of different quality. The differences seem to be related to uniformity and variation in the read counts. We have developed several measures to quantify different aspects of library quality:

- The total **number of sequenced reads** X, approximated by the sum of all read counts x_t:

$$X = \sum_{t=1}^{T} x_t \qquad \text{(eq. 2.10)}$$

- **Library complexity** C, here defined as the expected number of distinct molecules that could be obtained from infinitely deep sequencing. We approximate this complexity by fitting a non-linear least-squares (nls) model to a series of downsampled libraries. For each downsampled dataset, we obtain the number of distinct reads y and the number of total reads x (with duplicates). We fit the formula

$$y = C\frac{x}{k+x} \quad , \qquad \text{(eq. 2.11)}$$

  where k and C are fitted parameters. This model is a first-order approximation of the method implemented in `preseqR` by Daley and Smith [12]. We chose this first order approximation because the more complex `preseqR` approach failed to converge for too many of our single-cell libraries.

- The **spikiness** s of a library is a measure for the bin-to-bin variation of the read count x_t and is defined as:

$$s = \frac{\sum_{t=1}^{T-1} \left| x_{t+1} - x_t \right|}{\sum_{t=1}^{T} x_t} \qquad \text{(eq. 2.12)}$$

- By contrast, the **shannon entropy** *e* for the read count is a measure of the uniformity of the read distribution and is defined as:

$$e = -\sum_{t=1}^{T} \frac{x_t}{X} \cdot \log\left(\frac{x_t}{X}\right) \qquad \text{(eq. 2.13)}$$

  where X is the sum over all read counts: $X = \sum_{\tau=1}^{T} x_\tau$.

- We found that the **loglikelihood** of the model as determined from the Baum-Welch algorithm is also a good measure to discriminate libraries by quality. The higher the loglikelihood, the better the fit of the HMM.

- The **number of copy number segments** can also be used to assess library quality. A segment is defined as a continuous stretch of bins with the same copy number state. This number will be high for bad quality libraries and low for good quality libraries.

- The **Bhattacharyya distance** *b* is a measure of how well two distributions can be distinguished and is defined as:

$$b = -\log\left[\sum_{x} \sqrt{NB_1(r_1, p, x_t) \cdot NB_2(r_2, p, x_t)}\right] \qquad \text{(eq. 2.14)}$$

  where NB_1 is the negative bionomial distribution for the state monosomy and NB_2 is the negative bionomial distribution for the state disomy, and r and p are the dispersion and probability parameters thereof, respectively.

Please note that total number of reads, complexity, spikiness and Shannon entropy are defined on the read count, while loglikelihood, segment number, and Bhattacharyya distance are defined on the output of the Hidden Markov Model. While all of those measures allow quality assessment of single-cell libraries, we found that none of those measures alone was powerful enough to reliably discriminate good-quality and bad-quality libraries (see Figure 2-4 for examples). Therefore, we employed a multivariate clustering approach implemented in the R-package `mclust` [13], [14] to utilize all quality criteria simultaneously to discriminate libraries by quality. In general, for a good-quality library, number of reads, complexity, entropy, loglikelihood and Bhattacharyya distance should be high, while spikiness and number of segments should be low.

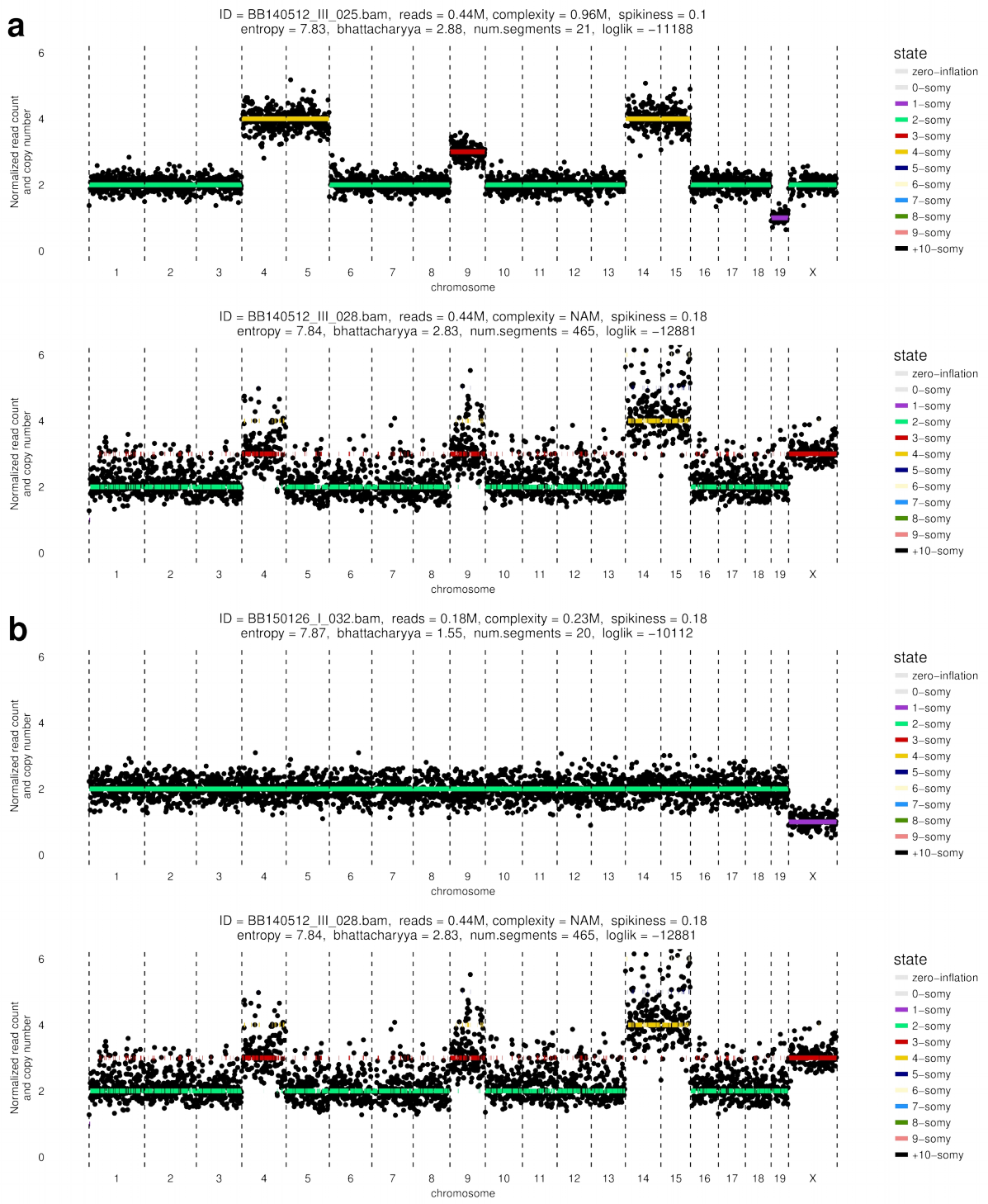


Figure 2-4 | Quality assessment of single-cell libraries. This figure illustrates the difficulties in assessing library quality if only one measure is available for quality assessment, *e.g.* (**a**) number of sequenced reads or (**b**) spikiness. Shown are copy number profiles with chromosomes on the x-axis and read count on the y-axis. Each dot represents a bin. **a** | Two cells with an identical number of sequenced reads (0.44M) and obviously very different library quality. However, the quality of these two cells can be distinguished by their spikiness. **b** | Two cells with identical spikiness (0.18) and obviously different library quality. However, the quality of these two cells can be distinguished by their number of segments or their loglikelihood.

Karyotype measures

To assess karyotype heterogeneity and the level of aneuploidy in populations of single cells we developed two measures that aggregate information over the population of single cells. For a set of N single cells with T bins, we define an aneuploidy score as:

$$D = \frac{1}{TN}\sum_{n=1}^{N}\sum_{t=1}^{T}|c_{n,t}-e_t| \qquad \text{(eq. 2.15)}$$

where $c_{n,t}$ is the copy number state of cell n at bin t, and e_t is the euploid copy number at bin t (*e.g.* $e = 2$ for autosomes, and $e = 2$ or 1 for the female or male X-chromosome respectively).

We define a heterogeneity score as:

$$H = \frac{1}{TN}\sum_{t=1}^{T}\sum_{f=0}^{S} f \cdot m_{f,t} \qquad \text{(eq. 2.16)}$$

where $m_{f,t}$ is the number of cells with copy number state s at bin t, and S is the total number of copy number states. Now, importantly, the $m_{f,t}$ is ordered for each bin such that $m_{f=0,t} \geq m_{f=1,t} \geq m_{f=2,t}$, etc. in such a way that f is not necessarily equal to s. Each type of aneuploidy (*e.g.* monosomy, trisomy, tetrasomy etc.) has an equal impact on this score, further evidenced in Table 2-1 by simulating various aneuploid conditions and calculating the aneuploidy and heterogeneity scores.

Table 2-1 | Simulating the effects of different ploidy-mixtures in a population on the aneuploidy and heterogeneity score. This table demonstrates that aneuploidy and heterogeneity score behave in an intuitive way. Specifically, each type of aneuploidy has an equal impact on the heterogeneity score, heterogeneity is maximal when ploidy-states are present in equal proportions, and both scores are independent of the total number of cells. (Source: Bakker and Taudt et al. 2016, [8])

#cells with ploidy in population	Aneuploidy	Heterogeneity
10 diploid	0	0
9 diploid + 1 monoploid	0.1	0.1
9 diploid + 1 triploid	0.1	0.1
9 diploid + 1 tetraploid	0.2	0.1
1 diploid + 9 monoploid	0.9	0.1
5 diploid + 5 monoploid	0.5	0.5
8 diploid + 2 triploid	0.2	0.2
8 diploid + 2 tetraploid	0.4	0.2
16 diploid + 4 tetraploid	0.4	0.2
8 diploid + 1 triploid + 1 tetraploid	0.3	0.3
7 diploid + 2 triploid + 1 tetraploid	0.4	0.4
7 diploid + 1 triploid + 1 tetraploid + 1 monoploid	0.4	0.6

Comparison to other methods

We compared `AneuFinder` with a web based tool for single-cell CNV calling based on Circular Binary Segmentation (`Ginkgo` [15]). We analysed 325 cells from tumors T158, T170, T257, T260, T386 and B-ALL-B with `Ginkgo` (1Mb variable-width bins based on simulated reads of 48 bp mapped with bowtie, otherwise default parameters) and `AneuFinder` at variable bin size 1Mb, and found that 81% (263/325) of the cells had concordant copy number calls for more than 95% of base pairs. Of the remaining 19% (62/325) of discordant cells, approximately half (27/62) were due to failed libraries. The other half (35/62) was caused by incorrect fits due to sequencing noise for `AneuFinder` (Figure 2-5|a) or wrongly chosen ploidy state for `Ginkgo` (Figure 2-5|b). However, the `AneuFinder` pipeline would filter out these problematic fits (like in Figure 2-5|a) in the quality control step and therefore these cells would not be used for subsequent analysis.

While `Ginkgo` was more robust to sequencing noise (Figure 2-5|a,c), we found it to be less sensitive for the detection of small CNVs of only a couple of bins (Figure 2-5|d). Another important advantage of `AneuFinder` is that it offers more flexibility for the analysis of non-standard genomes and sequencing parameters. While `Ginkgo` is a web-tool and very easy to use, `AneuFinder` is available as R-package and allows easy scripting of analyses workflows.

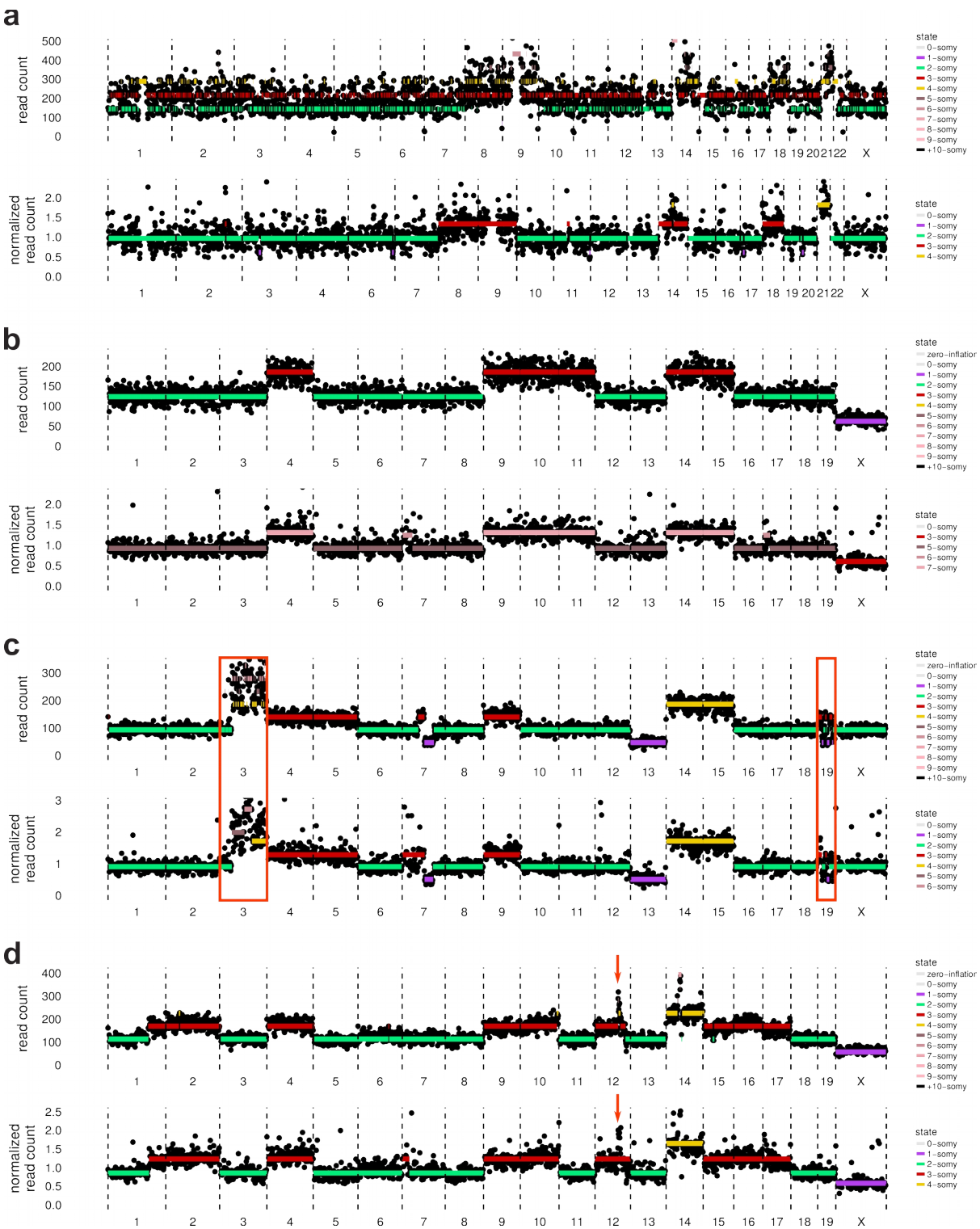


Figure 2-5 | Examples of discordant copy number calls between AneuFinder and Ginkgo. Top panels show the `AneuFinder` profiles, bottom panels show the `Ginkgo` profiles, respectively. **a** | Low quality library showing a highly segmented fit with `AneuFinder`. **b** | Wrongly chosen ploidy state with `Ginkgo`. **c** | Red boxes indicate chromosomes with unusually high read count dispersion where `AneuFinder` fails to assign a clear copy number state. **d** | Small copy number change that is detected with `AneuFinder` but not with `Ginkgo`. (Source: Bakker and Taudt et al. 2016, [8])

Discussion

AneuFinder is an implementation of a Hidden Markov Model for copy number calling from low-coverage scWGS data. The method works extremely well for detecting whole-chromosome aneuploidies and large sub-chromosomal copy number aberrations (CNA). The minimum size of a CNA that can be detected depends on the sequencing depth and quality of the scWGS experiment. A typical scWGS experiment [8], [9] yields a genomic coverage of around 1% or 0.01 X (up to 6% in [16]), meaning that 1% of all bases are covered by a sequencing read or alternatively that each base is covered on average by 0.01 reads. This is roughly equivalent to 0.6 million reads on a mouse genome. This is not much compared to other sequencing applications, where coverage can easily be 30 X (3000%), indicating that on average 30 reads overlap each base. A genomic coverage of 1% will therefore allow only relatively large CNA to be accurately called. Most of the analysis in section “Applications” (page 33) were performed with bin size 1 Mb, which roughly corresponds to 200 reads within each bin and allows reliable discrimination of copy number states. Figure 2-3|b shows the fitted distributions for a typical single cell with 0.3 million reads (mouse genome) that was analyzed at bin size 1 Mb. At this bin size and sequencing depth, the distributions for the different copy number states are nicely separable, but the separability would decrease with decreasing bin size or sequencing depth. We have not systematically investigated the dependence of resolution (chosen bin size) on sequencing depth or quality and this could be the subject of further research. Choosing an appropriate bin size for analysis remains a human task so far. It is worth noting that the bin size is the only parameter that a user needs to specify when using AneuFinder. There are other parameters which influence the convergence of the Baum-Welch algorithm, but these can be used with default values and will not require adjustment in most cases.

Another point of discussion is the confidence in the copy number calls. This seems to be relevant especially for small CNA, and could answer the question whether an observed small CNA is a real CNA or just noise. An option to assign such a confidence value for each bin would be the posterior probabilities which are the result of the Baum-Welch fitting procedure. These “posteriors” give the probability that a bin belongs to the reported copy number state, and a value close to 1 indicates high confidence in the predicted copy number state. However, these probabilities are currently not reported in AneuFinder. Another option to assess the validity of a small CNA would be the comparison with other single-cells: If all single-cells show the same CNA, then it would very likely not be noise but a real biological feature (or an artifact).

An important limitation of the presented HMM is that it is blind to whole-genome amplifications. For example, a fully tetraploid genome would be called diploid, because the HMM assumes that the most frequent copy number state is disomic. Distinguishing diploid from tetraploid cells (or any other ploidy) would require some sort of comparison between single-cells with the assumption that the amount of sequenced DNA is proportional to the DNA content of the cell. Whether this is the case and developing an algorithm to select the ground ploidy-state might be the goal of further research.

The semi-automated quality control is another point which could be further improved. Currently, a user performs quality clustering and selects good-quality clusters by hand for further analyses. This is a big improvement over selecting individual cells by hand, but a fully automated method would of course be preferable.

Also the HMM seems to be not ideal if the fitted distributions have too much overlap. This is the case when the bin size is too small, but also becomes a problem at high copy numbers (*e.g.* 9-somy and 10-somy), where the copy number state starts to oscillate between adjacent copy numbers (see Figure 2-5|a,c for such an effect due to high variance of the data).

Applications

Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies

Bjorn Bakker*, Aaron Taudt*, Mirjam E. Belderbos, David Porubsky, Diana C. J. Spierings, Tristan V. de Jong, Nancy Halsema, Hinke G. Kazemier, Karina Hoekstra-Wakker, Allan Bradley, Eveline S. J. M. de Bont, Anke van den Berg, Victor Guryev, Peter M. Lansdorp, Maria Colomé-Tatché and Floris Foijer

* these authors contributed equally

Contributions: Conception and implementation of computational analysis as stand-alone software package.

Genome Biology 2016; doi: 10.1186/s13059-016-0971-7

Background:

Chromosome instability leads to aneuploidy, a state in which cells have abnormal numbers of chromosomes, and is found in two out of three cancers. In a chromosomal instable p53 deficient mouse model with accelerated lymphomagenesis, we previously observed whole chromosome copy number changes affecting all lymphoma cells. This suggests that chromosome instability is somehow suppressed in the aneuploid lymphomas or that selection for frequently lost/gained chromosomes out-competes the CIN-imposed mis-segregation.

Results:

To distinguish between these explanations and to examine karyotype dynamics in chromosome instable lymphoma, we use a newly developed single-cell whole genome sequencing (scWGS) platform that provides a complete and unbiased overview of copy number variations (CNV) in individual cells. To analyse these scWGS data, we develop `AneuFinder`, which allows annotation of copy number changes in a fully automated fashion and quantification of CNV heterogeneity between cells. Single-cell sequencing and `AneuFinder` analysis reveals high levels of copy number heterogeneity in chromosome instability-driven murine T-cell lymphoma samples, indicating ongoing chromosome instability. Application of this technology to human B cell leukaemias reveals different levels of karyotype heterogeneity in these cancers.

Conclusion:

Our data show that even though aneuploid tumors select for particular and recurring chromosome combinations, single-cell analysis using `AneuFinder` reveals copy number heterogeneity. This suggests ongoing chromosome instability that other platforms fail to detect. As chromosome instability might drive tumor evolution, karyotype analysis using single-cell sequencing technology could become an essential tool for cancer treatment stratification.

Single-cell whole genome sequencing reveals no evidence for common aneuploidy in normal and Alzheimer's disease neurons

Hilda van den Bos, Diana C. J. Spierings, Aaron Taudt, Bjorn Bakker, David Porubský, Ester Falconer, Carolina Novoa, Nancy Halsema, Hinke G. Kazemier, Karina Hoekstra-Wakker, Victor Guryev, Wilfred F. A. den Dunnen, Floris Foijer, Maria Colomé Tatché, Hendrikus W. G. M. Boddeke and Peter M. Lansdorp

Contributions: Data analysis with AneuFinder.

Genome Biology 2016; doi: 10.1186/s13059-016-0976-2

Background:

Alzheimer's disease (AD) is a neurodegenerative disease of the brain and the most common form of dementia in the elderly. Aneuploidy, a state in which cells have an abnormal number of chromosomes, has been proposed to play a role in neurodegeneration in AD patients. Several studies using fluorescence in situ hybridization have shown that the brains of AD patients contain an increased number of aneuploid cells. However, because the reported rate of aneuploidy in neurons ranges widely, a more sensitive method is needed to establish a possible role of aneuploidy in AD pathology.

Results:

In the current study, we used a novel single-cell whole genome sequencing (scWGS) approach to assess aneuploidy in isolated neurons from the frontal cortex of normal control individuals (n = 6) and patients with AD (n = 10). The sensitivity and specificity of our method was shown by the presence of three copies of chromosome 21 in all analyzed neuronal nuclei of a Down's syndrome sample (n = 36). Very low levels of aneuploidy were found in the brains from control individuals (n = 589) and AD patients (n = 893). In contrast to other studies, we observe no selective gain of chromosomes 17 or 21 in neurons of AD patients.

Conclusion:

scWGS showed no evidence for common aneuploidy in normal and AD neurons. Therefore, our results do not support an important role for aneuploidy in neuronal cells in the pathogenesis of AD. This will need to be confirmed by future studies in larger cohorts.

Copy number alterations assessed at the single-cell level revealed mono- and polyclonal seeding patterns of distant metastasis in a small cell lung cancer patient

P. Ferronika, H. van den Bos, A. Taudt, D.C.J. Spierings, A. Saber, T.J.N. Hiltermann, K. Kok, David Porubsky, A.J. van der Wekken, W. Timens, F. Foijer, M. Colomé-Tatché, H.J.M. Groen, P.M. Lansdorp, and A. van den Berg

Contributions: Data analysis with AneuFinder. Bioinformatics analysis.

Annals of Oncology 2017; doi: 10.1093/annonc/mdx182

Intra-tumor heterogeneity (ITH) is a common feature of many cancers and can facilitate tumor evolution. In the present study we assessed intra-tumor copy number heterogeneity using low-coverage single-cell whole genome sequencing (scWGS). We determined copy number alterations (CNAs) in single cells of two areas of the primary tumor and from mediastinal lymph node, liver and adrenal metastases of a 79-year-old female stage IV small cell lung carcinoma (SCLC) patient. Copy number states in 2 Mb bins were assessed using a Hidden Markov Model in our custom developed pipeline `AneuFinder`. Of the 586 cells analysed, 346 passed quality control and were used for further analysis. Merged scWGS data of each tumor used to generate bulk CNA patterns were highly similar to those generated by array-based comparative genomic hybridization (aCGH) on DNA isolated from the same five tissue samples.

Unsupervised clustering of single-cell genomes for copy number similarities revealed a high degree of ITH among single cells from the primary tumor, lymph node and adrenal metastases, but a much lower degree of ITH with a distinct CNA pattern in the liver metastasis. The liver CNA pattern was characterized by a disomic, a trisomic and a tetrasomic part of chromosome 11, trisomy of 18, disomy of 22 and pentasomy of Xp.

Previous studies have shown that metastasis development may occur from single or multiple subclones of the primary tumor and from one metastatic site into another. To identify metastasis founder cells in this SCLC patient, we performed hierarchical clustering of all sequenced primary tumor and metastasis cells. This revealed an overall intermixed pattern, especially for the primary tumors, the lymph node and adrenal metastases. In contrast, the liver cells formed a distinct cluster that also contained two cells from primary tumor 1 and five adrenal metastasis cells, which all showed the liver-specific CNA pattern. The strong association of a subset of primary tumor and adrenal metastasis single cells to the merged liver metastasis CNA data was supported by higher Pearson's correlation coefficients to the merged liver as compared to their own merged CNA pattern. Together, these data indicate that in this SCLC patient the liver-metastasis founder cells were present in the primary tumor as a minor clone. Hierarchical clustering of all sequenced primary tumor and metastasis cells and the results of the Pearson's correlation coefficients showed close association of the liver metastasis cells, which supports the low ITH in liver metastasis.

In conclusion, we found a high degree of CNA heterogeneity among cells of five distinct tumor locations in a single SCLC patient. A minority of the tumor cells of the primary tumor and the adrenal gland metastasis showed the dominant CNA pattern observed in liver metastasis cells. Our data suggests polyclonal seeding of the lymph node and adrenal gland metastases and a monoclonal seeding of the liver metastasis in this patient.

Chapter 3

`AneuFinder2`: An algorithm for read-resolution copy number and breakpoint detection in single-cell whole genome and strand sequencing

Aaron Taudt[1,2], Diana C. J. Spierings[1], Peter M. Lansdorp[1,3], Maria Colomé-Tatché[1,2]

1. *European Research Institute for the Biology of Ageing, University of Groningen, University Medical Center Groningen, A. Deusinglaan 1, Groningen 9713 AV, The Netherlands.*
2. *Institute of Computational Biology, Helmholtz Zentrum München, Ingolstädter Landstr. 1, Neuherberg 85764, Germany.*
3. *Terry Fox Laboratory, BC Cancer Agency, Vancouver, BC V5Z 1L3, Canada. Division of Hematology, Department of Medicine, University of British Columbia, Vancouver, BC V6T 1Z4, Canada.*

Abstract

Strand-seq is a single-cell sequencing technique that allows the study of sister chromatid inheritance patterns, sister chromatid exchange (SCE) events, and identification of mis-oriented contigs in the reference genome. More recently, this technique has been applied to study sister chromatid exchange events in highly aneuploid cells with lots of copy number alterations (CNA). Existing tools where designed for diploid genomes and have difficulties in correctly assessing SCE events when copy number alterations are present. This chapter presents a computational method for mapping CNA and SCE from Strand-seq data. Copy number detection is based on a binary bisection and normalization method, and breakpoint (and SCE) detection is based on a statistical test with read-resolution. The method presented here can also be applied to single-cell whole genome sequencing (scWGS) data.

Introduction

During mitosis, DNA is replicated for subsequent segregation into daughter cells. This replication takes place at each chromosome (both paternal and maternal, which should be regarded as separate chromosomes for this explanation), and the two identical copies of a replicated chromosome are called sister chromatids. Faithful segregation of sister chromatids into the daughter cells is vital for the functioning of an organism. The dogma for many decades has been that segregation of sister chromatids into daughter cells is random, but other non-random mechanisms such as the Immortal Strand Hypothesis or Silent Sister Hypothesis have been suggested to explain experimental observations [17].

Strand-seq is a novel single-cell sequencing technique that allows the study of segregation patterns of sister chromatids *in vitro* and *in vivo* [18]. The idea behind this technique is to sequence only the template strand after replication in the daughter cells, and not the newly synthesized strand (see Figure 3-1). This is achieved by labeling the nascent strand with BrdU (bromodeoxyuridine), which is incorporated in the place of thymidine during cell replication and allows photolysis of the newly synthesized strand with UV treatment. The remaining template strand is then sequenced by Illumina sequencing. Strand-seq also allows the study of genomic rearrangements such as inversions or sister chromatid exchanges (SCE). SCE can occur before segregation when the two sister chromatids are still connected at the centromer.

Existing computational methods for the detection of breakpoints from Strand-seq data were designed with a diploid genome in mind (`BAIT` [19], `invertR` [20] and `breakpointR` [21]) and fail to accurately locate breakpoints when copy number alterations are present in the sequenced cell. Figure 3-1 shows the Strand-seq procedure for a triploid cell.

In this chapter I develop a method for the detection and mapping of breakpoints from Strand-seq data with arbitrary copy number profiles. More generally, this method is able to locate and classify any type of breakpoint, not only template strand switches, and can also be applied to locate copy number breakpoints (CNB) in single-cell whole genome sequencing (scWGS) data. The central idea is to call copy numbers for both the Watson and Crick strand (or both strands combined in case of scWGS) and define breakpoints as changes in copy number state. These breakpoints are then refined with read-resolution to make full use of the sequencing data.

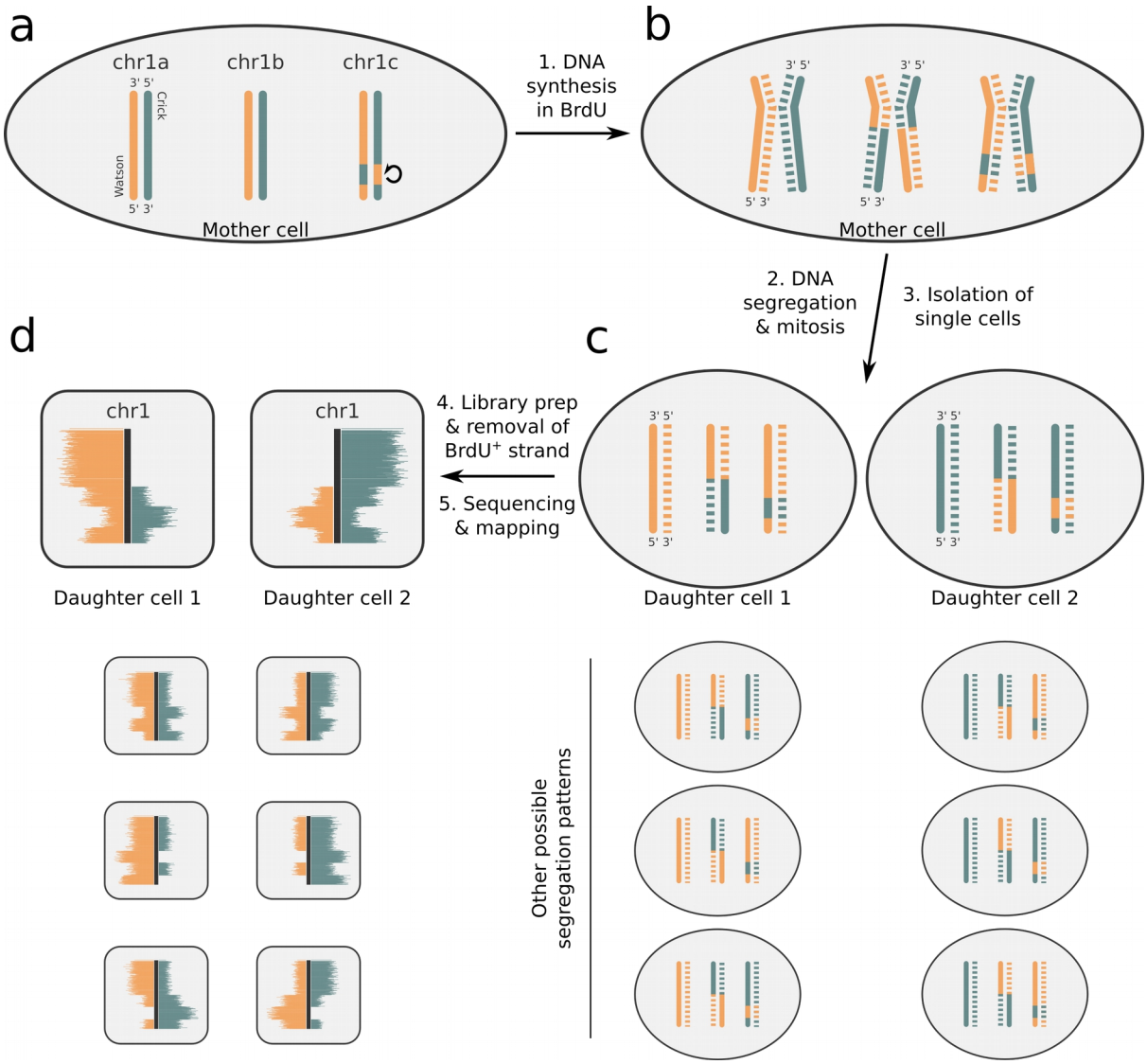


Figure 3-1 | Strand-seq for a triploid cell. **a** | Depicted are all three homologs of chromosome 1 of a triploid cell. The homologs are named chr1a, chr1b and chr1c, and could represent *e.g.* two copies of the maternal homolog and one copy of the paternal homolog. The Watson strand is colored orange, the complimentary Crick strand colored in teal. Chr1c shows an inversion. **b** | DNA replication in the presence of BrdU incorporates BrdU in the newly synthesized strand (dashed lines). Chr1b undergoes a sister chromatid exchange. **c** | After DNA segregation and mitosis, two triploid daughter cells each have their own set of three homologs. There are four possible segregation patterns with three homologs. **d** | UV light and Hoechst treatment are used for photolysis of the BrdU substituted strands (dashed lines), preventing them from subsequent PCR amplification and sequencing. As a result, only the original template strands (continuous lines) are sequenced and mapped to the reference genome. Ideograms show binned read counts along chromosome 1.

Model specification

Copy number detection with `AneuFinder2` (Figure 3-2) is similar to the approach described in Chapter 2, but with important differences. The main steps for copy number detection are: (1) Binning (same as in Chapter 2, page 19), (2) correction for GC content with a loess-fit (Figure 3-3), (3) copy number detection with a binary bisection and normalization method. This method has the advantage that it can call any number of copies, not only the ones specified in the Hidden Markov Model (HMM). In practice this method also seems to be more robust and is less sensitive to noise than the HMM from Chapter 2.

Additional steps are performed for breakpoint detection, namely (4) estimation of confidence intervals, (5) breakpoint refinement and (6) hotspot detection (Figure 3-2|b-c). The binning step is identical to the approach already described in Chapter 2 (page 19), and is hence not repeated here. The following sections describe steps 2-6, starting with GC correction.

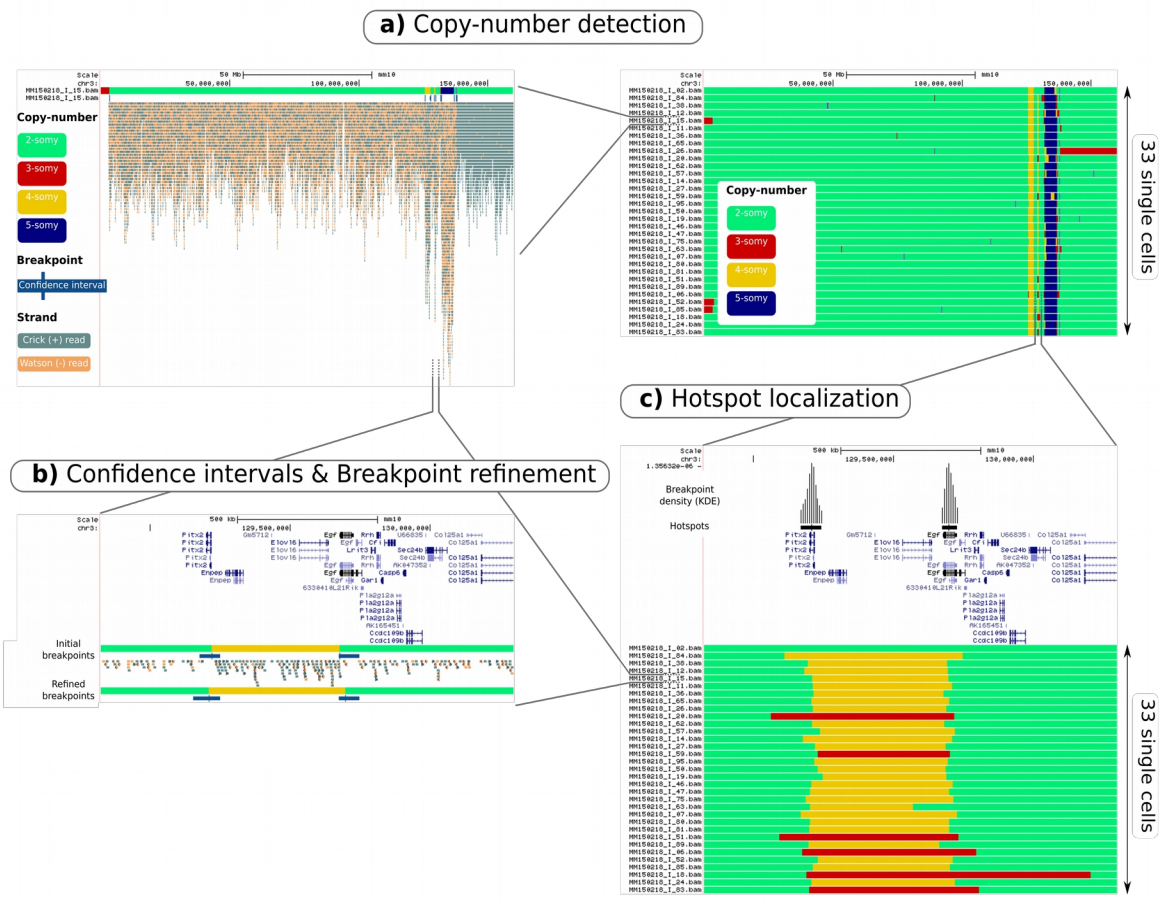


Figure 3-2 | AneuFinder2 – copy number and breakpoint detection in single-cell sequencing data. **a** | A UCSC genome browser snapshot of the copy number state of 33 single cells (right) and an extended view of one cell (left) showing also breakpoint locations and the sequenced read profile. **b** | Effect of the breakpoint refinement shown for a single cell and a small copy number variation on chromosome 3. **c** | Result of the hotspot localization showing the kernel density estimation (KDE) for significant hotspots in the 33 single cells.

GC-correction

Differing from Chapter 2, we have implemented a novel approach for GC correction which relies on a LOESS fit instead of the previously described quadratic fit. LOESS is a non-parametric regression method that fits the data by fitting simple models to localized subsets of the data. It does not require the specification of a function to fit the overall data, which makes it ideal for processes where no theoretical model exists.

For every bin t the GC-corrected read count x_t^{GC} is obtained by multiplying the read count x_t with a correction factor f_{GC} that is dependent on the GC content. This correction factor is determined by a loess-fit:

$$x_t^{GC} = x_t \cdot f_{GC} = x_t \cdot \frac{mean(x_t)}{loess(x \sim GC)_t} \quad , \qquad \text{(eq. 3.1)}$$

where mean(x_t) is the mean read count over all bins (trimmed by 5%). Intuitively, the read count in every bin is scaled in such a way that it is independent of GC content, while mean and variance of the read count distribution stay approximately the same (Figure 3-3).

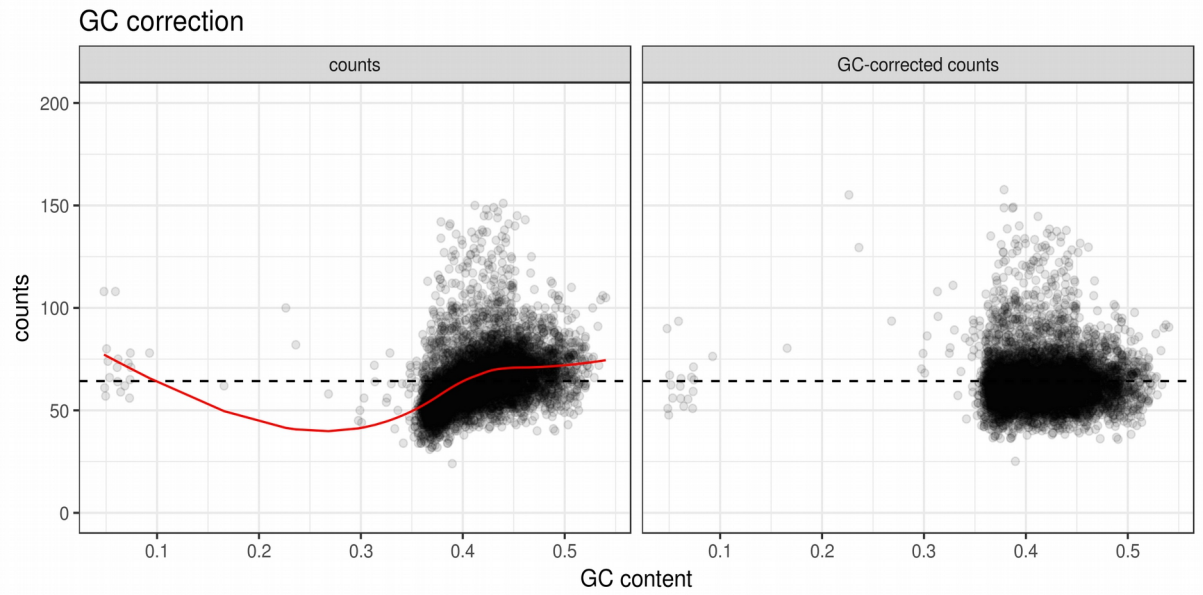


Figure 3-3 | GC correction for a typical single-cell. Raw read counts (left panel) show a GC-bias, which can be corrected for by the proposed correction procedure (right panel). The black dashed line indicates the mean read count mean(x_t), and the red line shows the loess-fit. (Library ID = MM150218_I_22.bam, binsize = 400 kb).

Copy number calling with changepoint algorithm

Strand-seq makes use of the separate mapping on the Watson and Crick strand, therefore, after binning, every bin contains a read count for both the Watson and Crick strand. Bins are segmented with the `e.divisive` function from the `ecp` package [22], which uses a binary bisection method and permutation test to obtain likely-correct copy number segments (Figure 3-4|a.1). The `ecp` package was designed to detect changepoints in any type of data

(not particularly genomic data), and is well suited to the task because it is able to use bivariate input (Watson and Crick reads) for the segmentation. The output is a segmentation into continuous-valued copy numbers. However, since we are analyzing single-cell data, we can assume that only discrete-valued copy numbers exist, and that the average read count of a segment with *N* copies is *N* times as high as the read count of a segment with only 1 copy. We employ an approach described in [15] to obtain discrete-valued copy number states (Figure 3-4|a.2). Briefly, read counts are normalized (such that the mean is at 1) and averaged within each segment (referred to as the raw copy number profile). These raw copy number counts *r* are then scaled with a factor *X* such that the sum-of-squares (SoS) error over all bins *t* is minimized:

$$X = \underset{X}{argmin}\left[\sum_{t}\left(r_t \cdot X - round\left(r_t \cdot X\right)\right)^2\right] \qquad \text{(eq. 3.2)}$$

The final copy number states are then calculated as

$$CN_t = round\left(r_t \cdot X\right) \ . \qquad \text{(eq. 3.3)}$$

After having determined the copy number profile, we assign count distributions to each state (Figure 3-4|a.3), which will later be used for the estimation of confidence intervals. For state 0-somy (0 copies), we assign a geometric distribution, where the mean is the (1% trimmed) mean of the read counts in state 0-somy. For the other states (1-somy, 2-somy, etc.) we assign a negative binomial distribution if the (1% trimmed) variance is bigger than the (1% trimmed) mean, a binomial distribution if the variance is smaller than the mean, and a poisson distribution if the variance is equal to the mean.

Estimation of confidence intervals

A breakpoint in the copy number profile is defined as the border of two bins where the copy number state changes. Therefore, the resolution of breakpoint detection is limited to the chosen bin size. To estimate confidence intervals with read-resolution (Figure 3-4|b), we go outwards from a breakpoint read by read (*e.g.* to the left) and test the probability that the reads within the tested interval belong on the other side of the breakpoint (*e.g.* on the right). Obviously, in doing so we calculate two confidence intervals, one for the left side and one for the right side of the breakpoint. A (left-sided) confidence interval with threshold $p = 0.01$ means that the probability is 1% that the reads within the confidence interval belong to the other (right) side of the breakpoint.

Another way to describe this is to say that for both left and right side of the breakpoint, we test the null hypotheses that the reads within the tested interval are generated by the distribution from the other side of the breakpoint.

Let x_{Side}^{Strand} be the number of reads within the tested interval on the respective *strand* (- or +) and *side* (left or right). The probability for the left side of the breakpoint is determined as follows:

$$p_{Left} = p_{Left}^{-} \cdot p_{Left}^{+} = CDF_{Right}^{-}\left(X \circ x_{Left}^{-}\right) \cdot CDF_{Right}^{+}\left(X \circ x_{Left}^{+}\right) \quad , \qquad \text{(eq. 3.4)}$$

where $CDF_{Side}^{Strand} = CDF\left(\mu_{Side}^{Strand}, \sigma_{Side}^{Strand}\right)$ is the cumulative density function of the assigned distribution (geometric for 0-somy, binomial or negative binomial for 1-somy, 2-somy, etc.), with μ and σ being the mean and variance. The ○-symbol indicates either ≤ or ≥ (*i.e.* if the lower or upper tail of the CDF is to be computed). The lower tail is computed if $\mu_{Left}^{Strand} < \mu_{Right}^{Strand}$, and the upper tail otherwise. This is to ensure that we compute the probability of finding the observed number of reads within the confidence interval *or a more extreme outcome*.

Similarly, the probability for the right side of the breakpoint is calculated as:

$$p_{Right} = p_{Right}^{-} \cdot p_{Right}^{+} = CDF_{Left}^{-}\left(X \circ x_{Right}^{-}\right) \cdot CDF_{Left}^{+}\left(X \circ x_{Right}^{+}\right) \quad . \qquad \text{(eq. 3.5)}$$

The lower tail of the CDF is computed if $\mu_{Right}^{Strand} < \mu_{Left}^{Strand}$, and the upper tail otherwise.

Importantly, mean and variance of the CDF need to be scaled to account for the length L_{Side} of the confidence interval. This scaling is achieved by multiplying mean and variance by L_{Side} / *binsize*.

Breakpoint refinement

The above described method is able to determine breakpoints with the resolution of the selected bin size. To further refine the breakpoints with read-resolution, the estimated confidence intervals are taken as an initial guess for the breakpoint location and the breakpoint is shifted to its most likely position within the confidence interval (Figure 3-4|c). The most likely position is determined as the position where the probability of finding the observed number of strand-specific reads to the left and right of the new breakpoint location is maximized, given the distributions of the adjacent copy number states. Let x_{Side}^{Strand} be the number of reads within the initial confidence interval on the respective *strand* (- or +) and *side* (left or right) of the new breakpoint location. The probability that is to be maximized is then:

$$p = \prod_{\substack{Strand \\ Side}} df_{Side}^{Strand}\left(x_{Side}^{Strand}, \mu_{Side}^{Strand}, \sigma_{Side}^{Strand}\right) \quad , \qquad \text{(eq. 3.6)}$$

where *df* is the assigned density function of the respective copy number state (geometric for 0-somy, binomial or negative binomial for 1-somy, 2-somy, etc.), with μ and σ being the mean and variance. Again, the mean and variance are scaled (multiplied) by L_{Side} / *binsize* to account for the length L_{Side} of the interval to the left and right of the new breakpoint location.

After the breakpoint locations have been refined, confidence intervals are re-estimated as described in the previous section.

Breakpoint-hotspot localization

Kernel density estimation (KDE) with a gaussian kernel is employed to generate a breakpoint-density profile along each chromosome (Figure 3-4|d). The bandwidth for KDE is set to the average distance between read fragments. To localize hotspots from the breakpoint-density profile, a p-value is calculated for each point in the profile by computing the likelihood of obtaining this density by random breakpoint placement. For each chromosome, random breakpoint placement with the observed number of breakpoints and subsequent KDE is repeated 100 times, and an empirical cumulative density function (eCDF) is calculated from the resulting density values. This eCDF is used to determine the p-value of finding the observed breakpoint-density. Multiple testing correction is done with the Benjamini-Yekutieli method. The default p-value for all breakpoint-hotspots is $p \leq 0.05$.

Hotspot detection with KDE might be improved by choosing a variable kernel that corresponds to the estimated breakpoint confidence interval. However, such an approach complicates the calculation of meaningful p-values. A single high-confidence breakpoint could already induce a hotspot, due to its localized high density. A solution would be to calculate the significance based on the integral over the breakpoint density. An implementation of this procedure has been tested but did not provide satisfactory and intuitively behaving p-values, and was therefore removed from the final implementation.

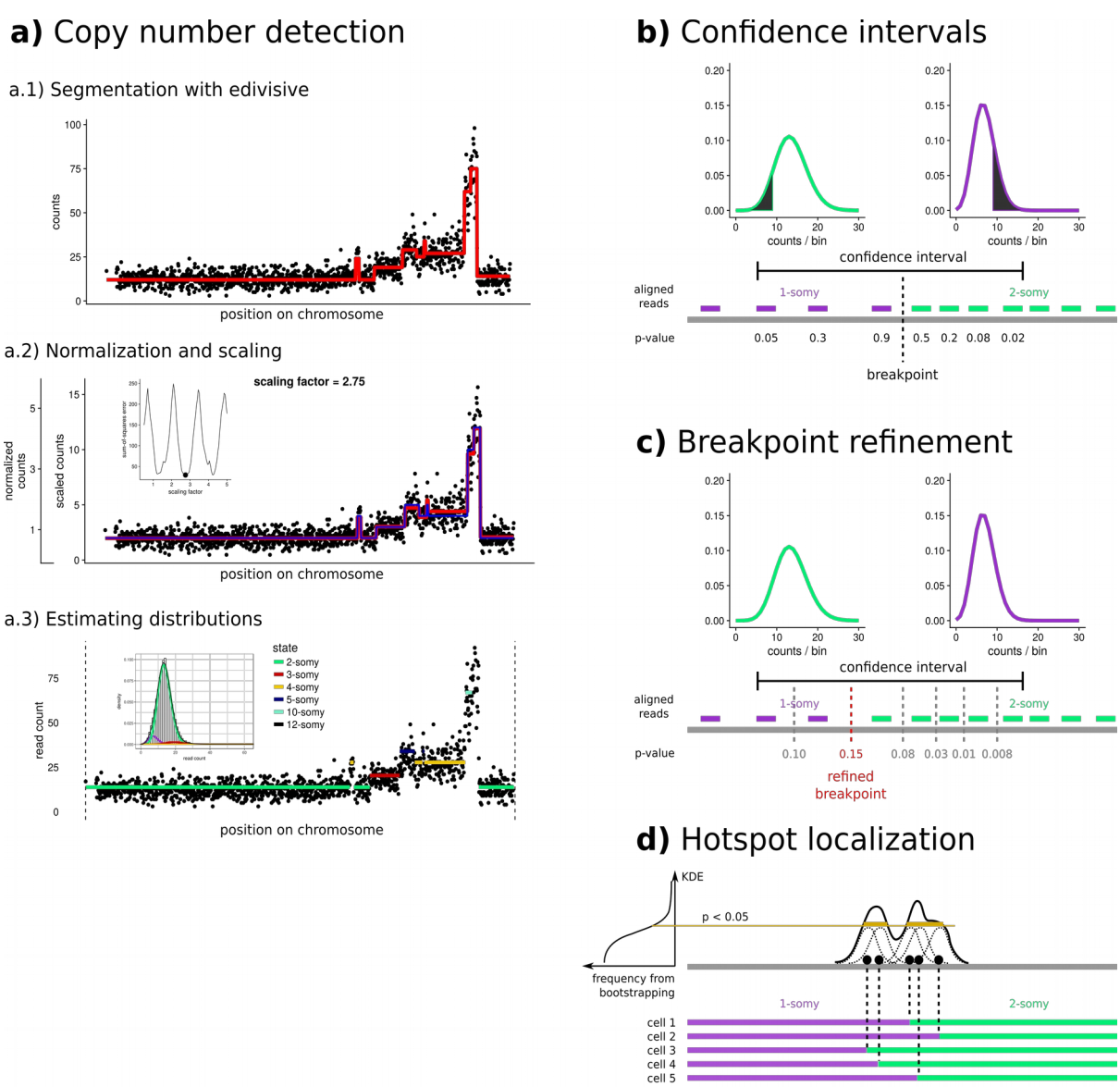


Figure 3-4 | AneuFinder2 – graphical explanation of the algorithm. For simplicity, the algorithm is shown for scWGS data with only one strand (both strands combined). For Strand-seq data, the procedure is similar but all steps are performed on Watson and Crick strands as described in the main text. **a | Copy number detection.** Graphs show binned read counts on chromosome 10 of a typical single cell. Each dot represents the read count in a 100 kb bin. (a.1) The first step is a segmentation on the binned read counts using the ecp package. (a.2) This is followed by a normalization such that the normalized counts are centered around 1, and a scaling procedure where the normalized counts are scaled to minimize the difference between the blue and red line. The red line shows the average (normalized and scaled) count in the respective segment, and the blue line is the copy number (= red line rounded to the nearest integer). (a.3) Distributions are estimated from the binned read counts for each copy number. **b | Confidence intervals.** For each breakpoint induced by the segmentation, a confidence interval is estimated by going outwards read by read and testing the probability that the observed number of reads is generated by the distribution from the other side of the breakpoint. **c | Breakpoint refinement.** Within the confidence interval, every possible location of the breakpoint is tested and the one that is most likely given the count distributions is picked. **d | Hotspot localization.** Breakpoint hotspots are localized with a kernel density estimation. Significance is determined with a bootstrapping approach.

Discussion

The algorithm for copy number calling presented in this chapter is an improvement over the HMM presented in the previous chapter. The resulting copy number states are less noisy compared to the HMM and do not oscillate in the case of badly separable distributions. This allows usage of smaller bin sizes, with the effect of increased sensitivity to small copy number alterations (CNA). This effect is quite substantial: While analyses with the HMM in Chapter 2 were usually conducted at bin size 1 Mb, analyses with the edivisive-approach can be performed at bin size 40 kb. This is 25 times smaller! A systematic comparison has not yet been performed though and might be the subject of further research.

The run time of the HMM from Chapter 2 is $O(N^2T)$, where N is the number of hidden states and T is the number of data points. The run time of `e.divisive` is $O(kT^2)$, where k is the number of estimated breakpoints. This means that the HMM scales linearly with the number of data points, while the run time of the edivisive-approach increases quadratically. Because the run time of `e.divisive` is also dependent on the number of estimated breakpoints, we expect that the edivisive-approach will become unusable for data sets with a large number of breakpoints per chromosome. However, this will not be the case for typical datasets.

The proposed procedure for breakpoint refinement relies on a good first guess, which in turn requires usage of a small bin size (*e.g.* 40 kb at 1% genomic coverage). Further research should determine the maximum bin size that is required to call accurate refined breakpoints with a given sequencing depth. Furthermore, it would also be interesting to see if the breakpoint refinement leads to the same results as a tiny bin size (*e.g.* 1 kb), which could serve as a validation for the refinement procedure. Such a tiny bin size would require a huge run time of at least several hours per cell.

Applications

Selective gene amplification in cultured organoid cells

Diana C. J. Spierings*, Martti Maimets*, Aaron Taudt*, Victor Guryev, Maria Colomé-Tatché, Robert P. Coppes and Peter M. Lansdorp

* these authors contributed equally

Contributions: Conception and implementation of computational analysis as stand-alone software package.

Manuscript in preparation

Embryonic or induced pluripotent stem cells are known to acquire individual changes next to recurrent changes that include genomic copy number alterations (CNAs) and dominant negative p53 mutations upon prolonged culturing. In contrast, adult organ-specific stem cells can be grown for years as organoids and these cells are reported to be remarkably stable, both genetically and phenotypically. However, using low-coverage single-cell DNA sequencing we show recurrent copy number gains at specific genomic locations in several independent long-term organoid cultures of murine salivary gland stem cells. Strikingly, these copy number gains are immediately adjacent to genes implicated in the EGF and Wnt/β−Catenin signaling pathways used to stimulate growth of these organoid cultures. Primary salivary gland stem cells, of which the cultures are derived from, reveal no recurrent CNAs. Our results support the power of single-cell DNA sequencing to detect local CNAs and suggest that cells in organoid cultures are not immune to the acquisition of genomic alterations which should be taken into account in future applications.

Chapter 4

chromstaR: An HMM for the combinatorial and differential analysis of ChIP-seq experiments

Aaron Taudt[1,2], Minh Anh Nguyen[3], Matthias Heinig[2], Frank Johannes[4], Maria Colomé-Tatché[1,2]

1. *European Research Institute for the Biology of Ageing, University of Groningen, University Medical Center Groningen, A. Deusinglaan 1, Groningen 9713 AV, The Netherlands.*
2. *Institute of Computational Biology, Helmholtz Zentrum München, Ingolstädter Landstr. 1, Neuherberg 85764, Germany.*
3. *Department of Radiation Oncology, The Netherlands Cancer Institute, Amsterdam, The Netherlands.*
4. *Department of Plant Sciences, Hans Eisenmann-Zentrum for Agricultural Sciences, Technical University Munich, Liesel-Beckmann-Str. 2, 85354 Freising, Germany.*

Adapted from Nature Reviews Genetics 2016; doi: 10.1038/nrg.2016.45
and bioRxiv 2016; doi: 10.1101/038612

Abstract

Post-translational modifications of histone residue tails are an important component of genome regulation. It is becoming increasingly clear that the combinatorial presence and absence of various modifications define discrete chromatin states which determine the functional properties of a locus. An emerging experimental goal is to track changes in chromatin state maps across different conditions, such as experimental treatments, cell-types or developmental time points. Here we present chromstaR, an algorithm for the computational inference of combinatorial chromatin state dynamics across an arbitrary number of conditions. chromstaR uses a multivariate Hidden Markov Model to determine the number of discrete combinatorial chromatin states using multiple ChIP-seq experiments as input and assigns every genomic region to a state based on the presence/absence of each modification in every condition. We demonstrate the advantages of chromstaR in the context of three common experimental data scenarios. First, we study how different histone modifications combine to form combinatorial chromatin states in a single tissue. Second, we infer genome-wide patterns of combinatorial state differences between two cell types or conditions. Finally, we study the dynamics of combinatorial chromatin states during tissue differentiation involving up to six differentiation points. Our findings reveal a striking sparcity in the combinatorial organization and temporal dynamics of chromatin state maps. chromstaR is a versatile computational tool that facilitates a deeper biological understanding of chromatin organization and dynamics. The algorithm is implemented as an R-package and freely available from http://bioconductor.org/packages/chromstaR.

Chromatin states – a review

DNA functionally interacts with a variety of epigenetic marks, such as cytosine methylation (5mC – 5-methylcytosine) or histone modifications (Figure 4-1|a). The dynamic placement of these marks along the genome is essential for coordinating gene expression programs and for maintaining genome integrity in response to developmental or environmental cues. Technological advances in the past decade have enabled high-resolution measurements of various epigenetic marks at a genome-wide scale [23], [24] (Figure 4-1|b). The computational integration of these measurements has led to the construction of so-called chromatin state maps (Figure 4-1|c), which provide an operational definition for the term "epigenome". These integrated maps are believed to give a good description of the functional state of the genome in a given cell type and at a specific time-point. Large initiatives are underway to collect reference epigenomes for different tissues, developmental stages, disease states and environmental treatments [25]–[27]. This information has already been instrumental in elucidating key chromatin changes during cellular differentiation, disease pathology and for functionally annotating causal variants from human genome-wide association mapping studies [27]–[29].

Chromatin state maps define epigenomes

Genomic DNA is tightly packed in cells, and the basic unit of DNA packaging is called the nucleosome. Approximately 150 bp of DNA wrap around a histone octamer, which consists of two copies of each of the core histones (H2A, H2B, H3 and H4). In addition to direct modifications of DNA in the form of 5mC, core histones can be subjected to a variety of chemical modifications of their amino acid residue tails [30] (Figure 4-1|a). Genome-wide maps of 5mC and various histone modifications can be readily obtained with array or next-generation sequencing (NGS) technologies coupled with bisulfite conversion or immunoprecipitation assays [23], [24] (Figure 4-1|b). To date, more than 100 histone modifications have been described [31], [32]. This large number has led to the idea of an epigenetic code [33], [34] – a layer of information that is encoded by recurring patterns of epigenetic marks. This code is potentially complex, as with 100 marks there are $2^{100} \approx 1.3 \times 10^{30}$ possible combinations of modifications at any given nucleosome. Although at a mechanistic level there are chemical restrictions on the co-occurrence of certain marks, the measured signal is an average over different cells and convoluted with noise, and spurious combinations may therefore be detectable. Nonetheless, integrative analysis of genome-wide maps based on a subset of all histone modifications have so far consistently revealed that only a small proportion of all possible combinations exist in the epigenome [27]–[29], [35]–[51] (Figure 4-2|a). This fact hints at strong biological restrictions on the placement of epigenetic marks. Despite this reduction in complexity, the inference of integrative chromatin states from individual array or NGS measurements continues to pose major computational and conceptual challenges that have not been fully solved. Chromatin states define a language that efficiently summarizes information across different marks and enables comprehensive comparisons across tissues, developmental stages and individuals. Large-

scale initiatives have made extensive use of those definitions and have produced reference epigenomes for various cell types and conditions in model and non-model species. Comparisons of these reference epigenomes have provided several insights into epigenomic variation. A major insight is that chromatin states corresponding to enhancer elements are most variable between tissue types [27], [40] and developmental time points [51], [52], whereas chromatin signatures corresponding to transcribed regions, transcription start sites (TSSs) or repressed regions are less variable [27]. Certain elements termed cREDS (cis-regulatory elements with dynamic signatures) are found with a strong promoter signature in one tissue but with an enhancer signature in other tissue types [27], [53], thus blurring the distinction between enhancer and promoter sequence elements [54].

Definitions of chromatin states

Since the proposition of the existence of a "histone code" in 2000 [33], [34], considerable effort has been spent to decipher this code, and many computational approaches have been developed to integrate single marks into chromatin state maps. Different conceptual ideas of a chromatin state underlie the different approaches. The original notion of a histone code [33], [34] is based on a molecular view that assumes that histone modifications (or epigenetic marks in general) are either present or absent at any given position in the genome in a binary manner, so that their combined presence and absence patterns define distinct combinatorial chromatin states [35], [44], [48], [51], [55]. A second view takes into account the continuous nature of the ChIP–seq signal and defines chromatin signatures on the basis of the signal shape rather than by the binary presence or absence of every mark [36], [56]. A third view defines probabilistic chromatin states [28], [39], [40], [43], [46] (also called fuzzy chromatin states [48]), which have probabilities associated with finding each mark in a given state, meaning that one state can be a superposition of multiple combinatorial patterns with different probabilities. A fundamental problem of chromatin-state-calling algorithms is to infer the 'true' number of states. Although it is reasonable to assume that the number of states increases with the number of epigenetic marks, our review of the literature shows that there is no clear trend (Figure 4-2|a). There are several reasons for this: first, different experimental techniques and analytical approaches investigate the epigenome at different resolutions, with higher resolution potentially leading to more chromatin states; second, the number of chromatin states is a function of the investigated marks (a set of uncorrelated marks has more states than a set of correlated or redundant marks); third, the majority of computational methods treat the number of chromatin states as an input rather than an output of the analysis, so that chromatin states reflect previous knowledge of chromatin. Another interesting question is the percentage of the genome that is covered with epigenetic marks or, conversely, is devoid of any marks. Our review of the literature shows that the percentage of empty epigenome decreases when more marks are measured (Figure 4-2|b). Indeed, one experiment involving 53 marks by Filion et al. [43] showed that essentially no part of the genome is permanently without epigenetic modifications.

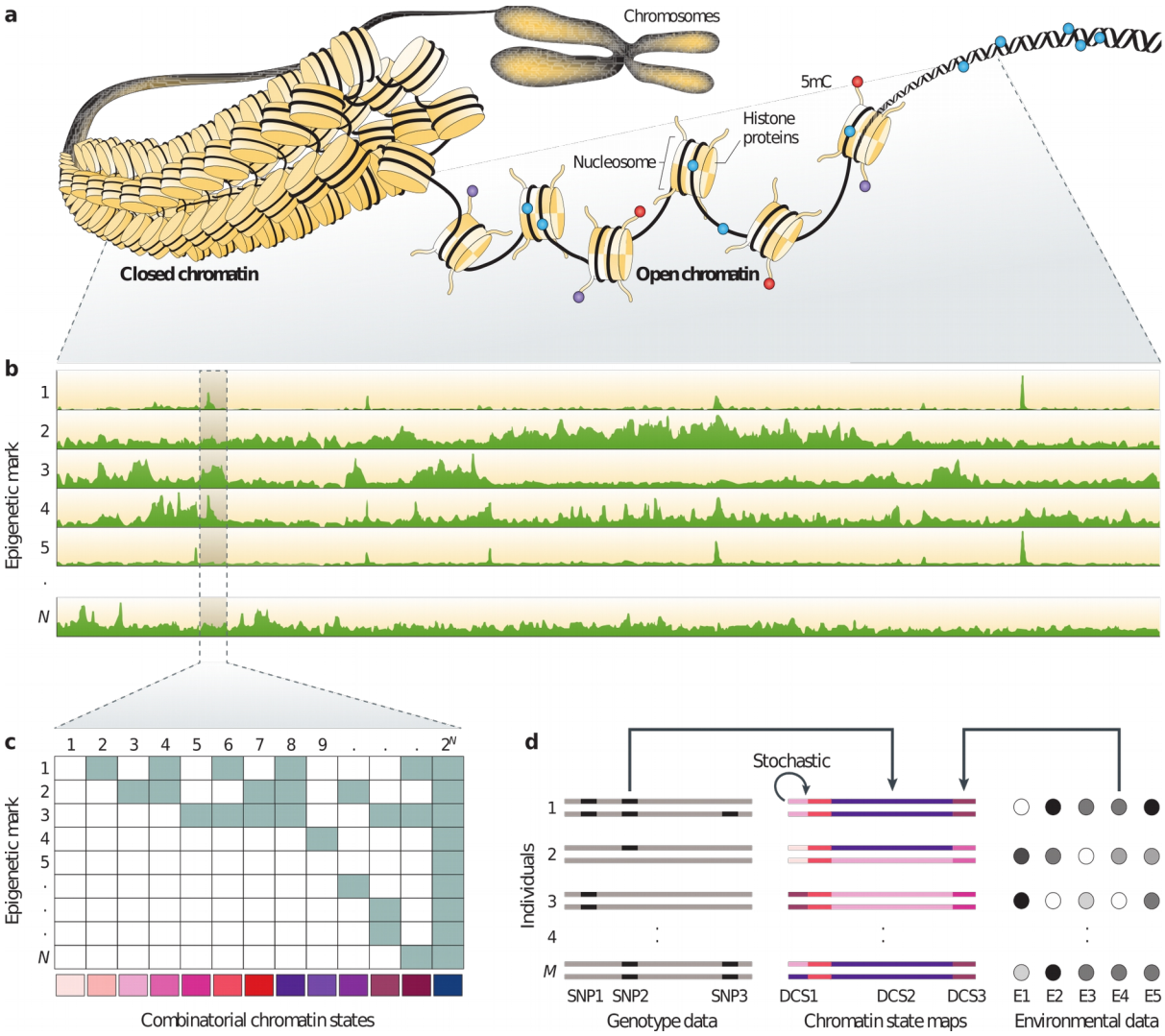


Figure 4-1 | Main steps in population epigenomic analysis. **a** | DNA is tightly packaged in cells and is functionally modified by a variety of epigenetic marks, such as cytosine methylation (5mC) or post-translational changes in histone proteins. The co-occurrence of specific epigenetic marks in a genomic region defines its functional state. Of note, histones in closed chromatin also contain repressive marks (not shown). **b** | The genome-wide distribution of different epigenetic marks can be measured using next-generation sequencing (NGS) technologies. Shown are the read-tracks from NGS measurements of *N* different epigenetic marks along the genome. **c** | The computational challenge is to infer distinct chromatin states for each genomic position. These chromatin states are defined by the joint presence and absence patterns of the different epigenetic marks. With *N* marks there can be 2^N possible combinatorial states. The color code on the bottom denotes each unique state. This analysis leads to the construction of chromatin state maps. **d** | Shown are the chromatin state maps of *M* diploid individuals. Individuals differ in their chromatin states in three genomic regions. These differential chromatin states (DCSs) can originate from DNA sequence polymorphisms, environmental factors or from stochastic changes. DCS2 is caused by a single-nucleotide polymorphism (SNP2), DCS3 is caused by exposure to environmental factor E4 and DCS1 is the result of stochastic processes in the mitotic maintenance of the chromatin state at that locus. The statistical challenge is to try to identify these causal factors from millions of measured SNPs and a large number of environmental factors. (Source: Taudt et al. 2016, [57])

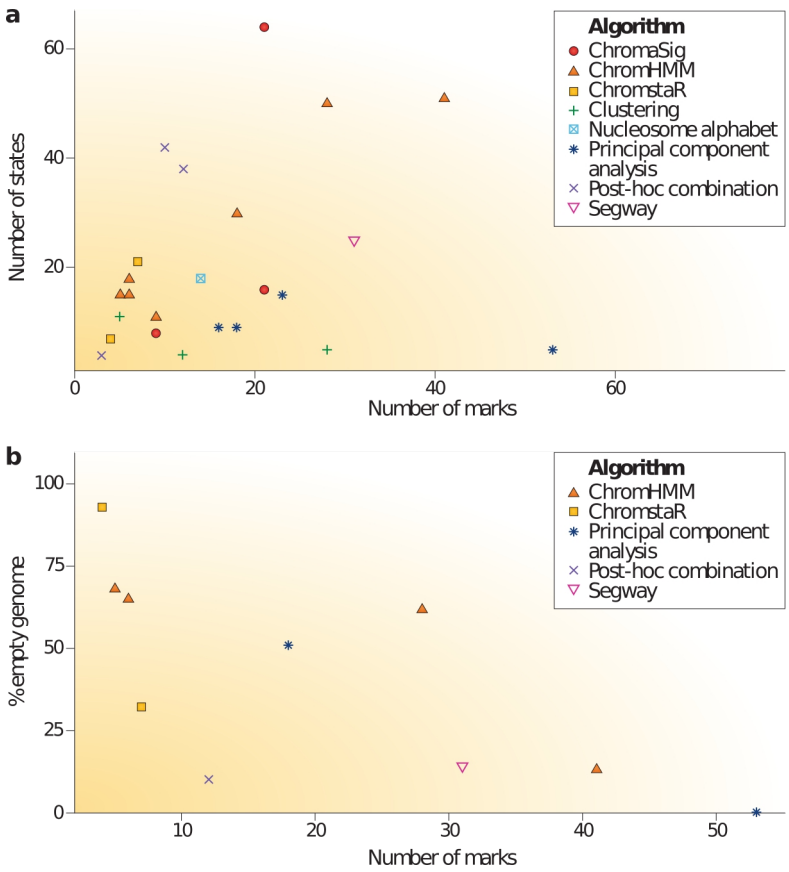


Figure 4-2 | Chromatin states in the literature. **a** | The number of chromatin states increases with the number of marks analyzed. The studies consistently show that only a small fraction of all possible chromatin states is present in the genome. **b** | The percentage of empty epigenome decreases with the number of marks analyzed, suggesting that no part of the genome is permanently without histone modification. (Source: Taudt et al. 2016, [57])

Introduction to chromstaR

Epigenetic marks such as DNA methylation or histone modifications play a central role in genome regulation. They are involved in a diversity of biological processes such as lineage commitment during development [58], maintenance of cellular identity [59], [60] and silencing of transposable elements [61]. The modification status of many histone marks has been extensively studied in recent years, first with ChIP-chip and later with ChIP-seq, now the de-facto standard procedure for genome-wide mapping of protein-DNA interactions and histone modifications. Since its advent in 2007 [58], [59], [62], ChIP-seq technologies have been widely used to survey genome-wide patterns of histone modifications in a variety of organisms [59], [63], [64], cell lines [65] and tissues [26], [27]. The multitude of possible histone modifications has led to the idea of a "histone code" [33], [34], a layer of epigenetic information that is encoded by combinatorial patterns of histone modification states (Figure 4-3|a). Major resources have been allocated in recent years to decipher this code, culminating in projects such as the ENCODE [66] and Epigenomics Roadmap [27]. Following their examples, most experiments nowadays are designed to probe several histone modifications at once, and often in various cell types, strains and at different developmental time points. These types of experiments pose new computational challenges, since initial solutions were designed to analyze one modification and condition at a time, therefore treating them as independent. Indeed, a commonly used strategy has been to perform peak calling for each experiment separately (univariate analysis) and to combine the peak calls post-hoc into combinatorial patterns [37], [48]. This approach is problematic for several reasons: Because of the noise associated with ChIP-seq experiments and peak calling, combining univariate peak calls will lead to the discovery of spurious combinatorial states that do not actually occur in the genome. Furthermore, different tools or parameter settings are often used for different modifications (*e.g.* peak calling for broad or narrow marks), making the outcome sensitive to parameter changes and control of the overall false discovery rate difficult. Lastly, this approach requires ample time and bioinformatic expertise, rendering it impractical for many experimental scientists.

Accurate inferences regarding combinatorial histone modification patterns are necessary to be able to understand the basic principles of chromatin organization and its role in determining gene expression programs. One way forward is to develop computational algorithms that can analyze all measured histone modifications at once (*i.e.* combinatorial analysis) and across different conditions (*i.e.* differential analysis). Some methods have been designed to integrate histone modifications into unified chromatin maps [36], [46], [55], [56], [67]–[71]. These methods can be classified into three different categories [57]:

- combinatorial, which define chromatin states based on the presence/absence of every histone modification [55],
- continuous, which define chromatin states based on the shape of the ChIP-seq signal [36], [56], and

- probabilistic, which have probabilities associated with finding each mark in a given state [46], [67]–[71].

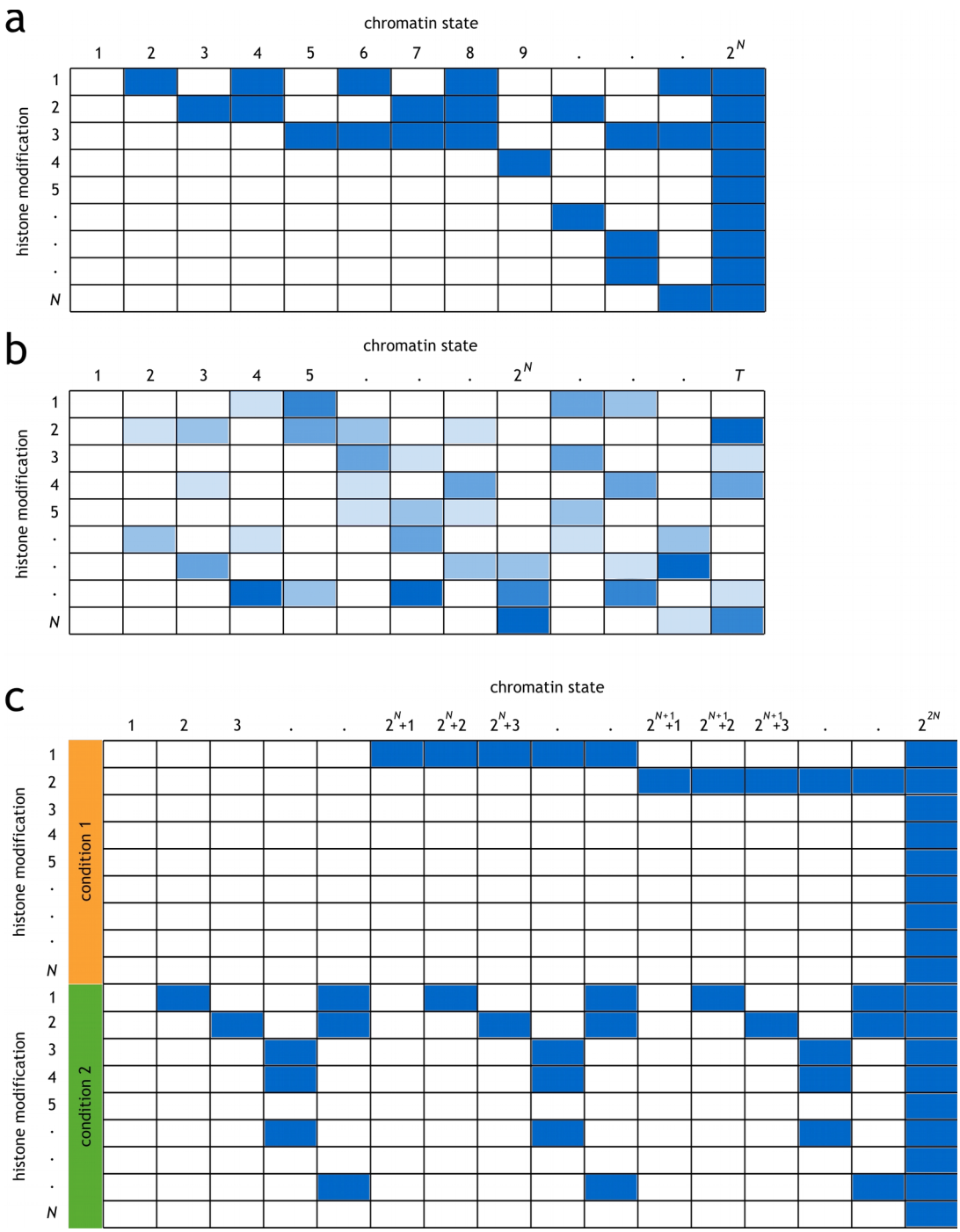


Figure 4-3 | Definition of chromatin states. **a** | Combinatorial chromatin state definition: Based on the presence (blue) or absence (white) of a histone modification, a chromatin state is the combination of the presence/absence calls at a given position. With *N* histone modifications there are 2^N different chromatin states. **b** | Probabilistic chromatin state definition: Each chromatin state has a probability (shades of blue) of finding a histone modification at a given position. Note that a probabilistic state can consist of multiple combinatorial states and vice versa. There is in principle no upper limit for the number of possible probabilistic chromatin states (here, *T*). **c** | Differential combinatorial chromatin states across two conditions: Based on the presence (blue) or absence (white) of a histone modification across different conditions. With *N* histone modifications and *M* conditions there are $2^{N \times M}$ different states. (Source: Taudt et al. 2016, [51])

A major drawback of the majority of these approaches [46], [67]–[71] is the need to specify the number of distinct chromatin states beforehand, which is usually not known *a priori*. Moreover, in the probabilistic interpretation the inferred states can consist of multiple and overlapping combinatorial states (Figure 4-3|b). This probabilistic state definition is useful to reduce noise and to identify functionally similar genomic regions for the purpose of annotation, but at the same time obscures a more direct interpretation of combinatorial states in terms of the presence/absence patterns of the underlying histone modifications.

Finally, none of these methods is designed for comparing chromatin maps across conditions. `ChromDiff` [72] and `dPCA` [73] are comparative methods that identify significant chromatin differences between groups of samples. `ChromDiff` discovers groups of epigenomic features which are the most discriminative in group-wise comparisons of samples, while `dPCA` uses a small number of differential principle components to explore differential chromatin patterns between two groups of samples. Both methods are useful for identifying defining features of each group, however they do not provide complete information about the genome-wide localization of all chromatin differences between all samples.

In order to overcome these problems we have developed `chromstaR`, a method for multivariate peak- and broad-region calling. `chromstaR` has the following conceptual advantages:
1) Every genomic region is assigned to a discrete, readily interpretable combinatorial chromatin state, based on presence/absence of every histone mark, providing a mechanistic interpretation of chromatin states which allows for better insights into how they regulate genome function. 2) The number of chromatin states does not have to be preselected but is a result of the analysis. 3) Histone modifications with narrow and broad profiles can be combined in a joint analysis along with an arbitrary number of conditions. 4) The same approach can be used for mapping combinatorial chromatin states in one condition, or for identifying differentially enriched regions between several conditions, or for both situations combined. 5) Our formalism offers an elegant way to include replicates as separate experiments without prior merging.

We demonstrate the advantages of `chromstaR` in the context of three common experimental scenarios (Figure 4-4|b). First, we consider that several histone modifications have been collected on a single tissue at a given time point (Figure 4-4|b, Application 1). The goal is to infer how these different modifications combine to form distinct combinatorial chromatin states and to describe their genome-wide distribution. Second, we consider that several histone modifications have been collected in two different cell types or conditions (Figure 4-4|b, Application 2). Here, the goal is to infer genome-wide patterns of combinatorial state differences between cell types or conditions. Third, we consider the more complex scenario where several histone modifications have been collected for multiple different time points or tissue types (Figure 4-4|b, Application 3). In this case, the goal is to infer how combinatorial chromatin states are modified during tissue differentiation or development. These three experimental scenarios broadly summarize many of the data problems that biologists and

bioinformaticians currently face when analyzing epigenomic data. We show that `chromstaR` provides biologically meaningful results to these types of data problems, and facilitates deeper biological insights into the dynamic coordination of combinatorial chromatin states in genome regulation.

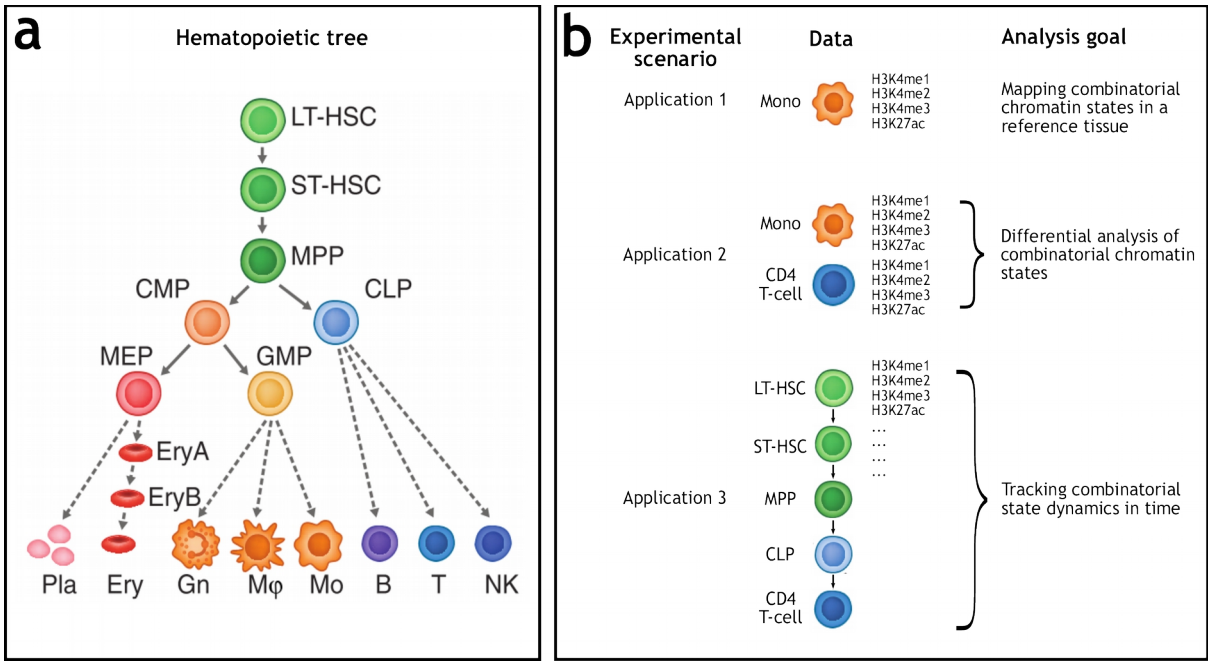


Figure 4-4 | Hematopoietic tree. **a** | Hematopoietic tree with cells that were probed by Lara-Astiaso et al. [52]. **b** | Three experimental scenarios that are investigated in this manuscript. (Source: Taudt et al. 2016, [51])

Model specification

Conceptual overview

Consider *N* ChIP-seq experiments: *N* histone modifications measured in one condition, or one histone modification measured in *N* conditions, or a combination of the two. After mapping the sequencing reads to the reference genome our method consists of two parts (Figure 4-5), a univariate peak calling step to estimate the distribution parameters, and a multivariate peak calling step to integrate information from all experiments:

(1) Univariate peak calling (Figure 4-5|a): For each ChIP-seq experiment, we partition the genome into non-overlapping bins (default 1 kb) and count the number of reads that map into each bin (*i.e.* the read count) [11]. We model the read count distribution as a two-component mixture of zero-inflated negative binomials [74], [75], with one component at low number of reads that describes the background noise and one component at high number of reads describing the signal. We use a univariate Hidden Markov Model (HMM) with two hidden states (*i.e.* unmodified, modified) to fit the parameters of these distributions [76].

(2) Multivariate peak calling (Figure 4-5|b): We consider all ChIP-seq experiments at once and assume that the multivariate vector of read counts is described by a multivariate distribution which is a mixture of 2^N components. We use a multivariate HMM to assign every bin in the genome to one of the multivariate components. The multivariate emission densities of the multivariate HMM, with marginals equal to the univariate distributions from step (1), are defined using a Gaussian copula [77].

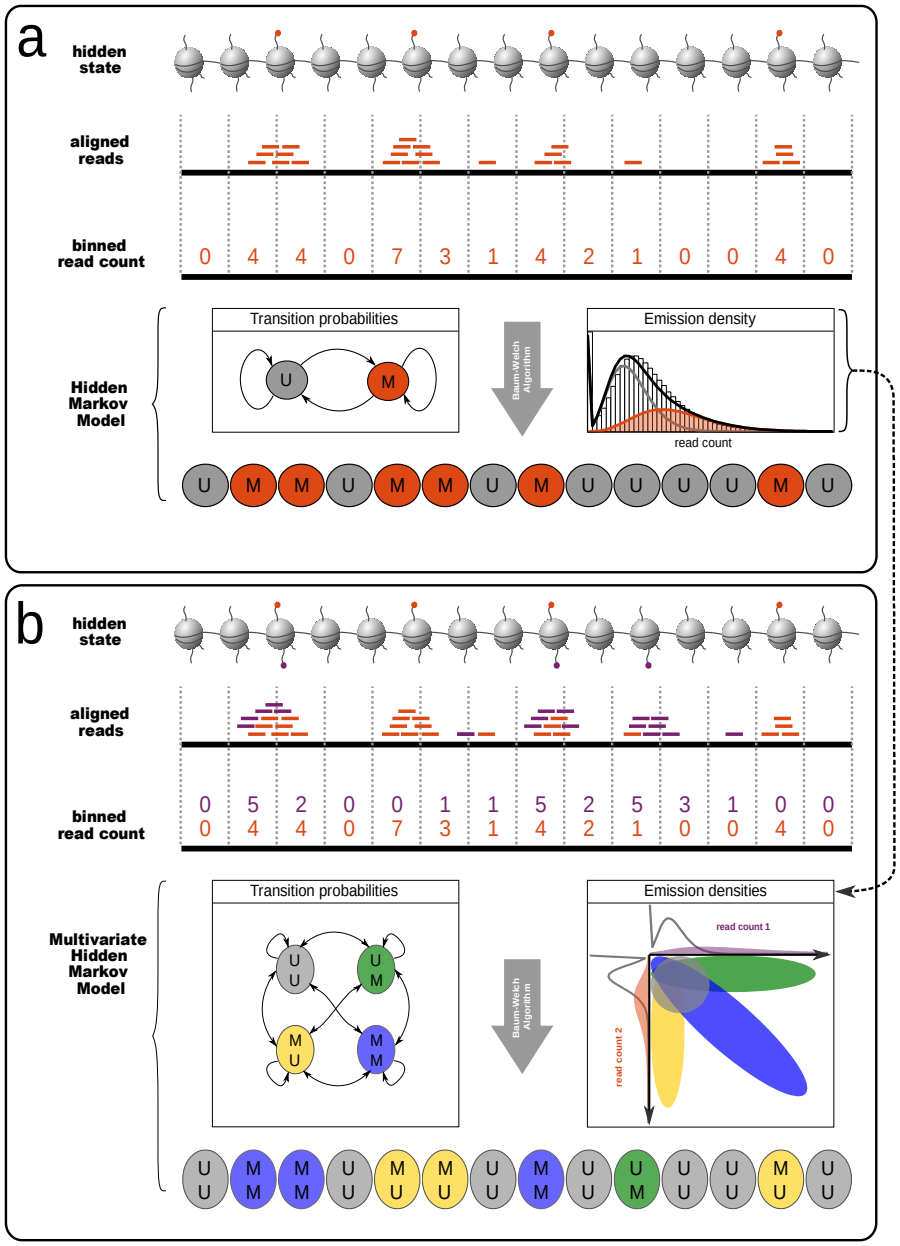


Figure 4-5 | Overview of analytical approach. **a** | Univariate peak calling: Aligned reads are counted in equidistant, non-overlapping bins. The resulting read count serves as observable for a Hidden Markov Model with components "unmodified" (U) and "modified" (M). Parameters are fitted with the Baum-Welch algorithm. **b** | Multivariate peak calling: A multivariate emission density is constructed from the densities of step (**a**), illustrated here for two dimensions. With N ChIP-seq experiments, the resulting multivariate density has 2^N components. For the two illustrated dimensions (ChIP-seq 1 and 2), univariate densities are plotted on the x- and y-axis, and the 4 components of the multivariate density are indicated by shaded areas, corresponding to both unmodified (UU – gray), both modified (MM – blue) and unmodified/modified (MU – yellow, UM – green). (Source: Taudt et al. 2016, [51])

Univariate Hidden Markov Model

For each individual ChIP-seq sample, we partition the genome into T non-overlapping, equally sized bins (default 1 kb). We count the number of aligned reads (regardless of strand) that overlap any given bin t and denote this read count with x_t. Following others [74], [75], we model the distribution of the read counts x with a two-component mixture of (zero-inflated) negative binomial distributions. In our case, the first component describes the *unmodified* regions and is modeled by a zero-inflated negative binomial distribution. The second component describes the *modified* regions and is modeled by a negative binomial distribution. Furthermore, for computational efficiency, we split the first component into the zero-inflation and the negative binomial distribution [76]. Our univariate Hidden Markov Model has thus three states i: *zero-inflation* (ZI), *unmodified* (U) and *modified* (M). We write the probability of observing a given read count x_t as

$$P(x_t|\theta) = \gamma_{ZI,t} B_{ZI}(x_t) + \gamma_{U,t} B_U(x_t|r_U, p_U) + \gamma_{M,t} B_M(x_t|r_M, p_M) \qquad \text{(eq. 4.1)}$$

where γ_{it} are the mixing weights (posterior probabilities) and B_i the emission distributions. The emission distribution of state *zero-inflation* is defined as

$$B_{ZI}(x_t) = \begin{pmatrix} 1, & \text{if } x_t = 0 \\ 0, & \text{if } x_t > 0 \end{pmatrix} \qquad \text{(eq. 4.2)}$$

and the emission distributions of state *unmodified* and *modified* are defined as negative binomial distributions

$$B_i(x_t|\,\theta_i = (r_i, p_i)) = \frac{\Gamma(r_i + x_t)}{\Gamma(r_i) x_t!} p_i^{r_i} (1-p_i)^{x_t} \qquad \text{(eq. 4.3)}$$

where Γ denotes the Gamma function and p and r denote the probability and dispersion parameter of the negative binomial distribution, respectively. Model parameters are fitted with the Baum-Welch algorithm [1]. The derivation of the updating formulas is detailed below, and uses notation introduced in [6]. Please see section “Mathematical notation“ in the introduction for details about the notation.

The conditional expectation Q that needs to be maximized can be written as

$$Q = \sum_i^N \gamma_{i,t=0} \log(\pi_i) + \sum_{i,j,t}^{N,N,T-1} \xi_{ijt} \log(A_{ij,c_{t,t+1}}) + \sum_{i,t}^{N,T} \gamma_{it} \log(B_{it}) \quad . \text{(eq. 4.4)}$$

The updated parameters for the negative binomial distributions can be obtained by solving $\frac{\partial Q}{\partial p} = 0$ and $\frac{\partial Q}{\partial r} = 0$. For independent negative binomial distributions, this yields

$$p_i = \left(\sum_t^T \gamma_{it} \cdot r_i \right) / \left(\sum_t^T \gamma_{it} \cdot (r_i + x_t) \right) \quad \text{and} \qquad \text{(eq. 4.5)}$$

$$\frac{\partial Q}{\partial r_i} = \sum_t^T \gamma_{it} \cdot \left(\log(p_i) - \Psi(r_i) + \Psi(r_i + x_t) \right) = 0 \quad , \qquad \text{(eq. 4.6)}$$

where $\Psi(x) = \frac{\partial}{\partial x} \log \Gamma(x)$ is the digamma function. The equation for r_i cannot be solved analytically, but can be solved with a numerical Newton-Raphson method to obtain the updated parameters.

Finally, we call a bin *modified* if the posterior probability $\gamma_{M,t} > 0.5$ and *unmodified* otherwise.

Multivariate Hidden Markov Model

Given N individual ChIP-sep samples with states *unmodified* (U) and *modified* (M), the number of possible combinatorial states is 2^N. Let $\boldsymbol{x}_t$ be the vector of N read counts for the t-th bin. The probability of observing a random vector $\boldsymbol{x}_t$ can be written as a mixture distribution of 2^N components:

$$P(\boldsymbol{x}_t) = \sum_{i=1}^{2^N} \gamma_{it} C_{it} \quad . \qquad \text{(eq. 4.7)}$$

Again, the γ_{it} denote the mixing weights and C_{it} denote the value of the emission distribution at bin t. We assume that the marginal densities of the multivariate count distributions C_i are given by the univariate distributions B described in the previous section. A convenient way to construct a multivariate distribution from known marginal (univariate) distributions is copula theory [77], [78]. Under the assumption of a Gaussian copula, the multivariate emission density C_i for combinatorial state i can be written as

$$C_i(\boldsymbol{x}_t) = \prod_{j=1}^{N} B_{ij}(x_{jt}) \cdot |\boldsymbol{\Sigma}_i|^{-1/2} \exp\left(-\frac{1}{2} \boldsymbol{z}_{it} (\boldsymbol{\Sigma}^{-1} - \boldsymbol{I}) \boldsymbol{z}_{it}^T \right) \quad , \qquad \text{(eq. 4.8)}$$

$$\text{with} \quad \boldsymbol{z}_{it} = \left(\phi^{-1}[F_{i,1}(x_{1,t})],\ \phi^{-1}[F_{i,2}(x_{2,t})],\ \ldots,\ \phi^{-1}[F_{i,N}(x_{N,t})] \right) \quad , \qquad \text{(eq. 4.9)}$$

where B_{ij} are the marginal density functions for combinatorial state i and $\boldsymbol{\Sigma}_i$ is the correlation matrix between the transformed read counts $\boldsymbol{z}_{it} = \varphi^{-1}[F_i(\boldsymbol{x}_t)]$. The cumulative distribution function (CDF) of B_{ij} is denoted by F_{ij}, while φ^{-1} denotes the inverse of the CDF of a standard normal [79].

The correlation matrix $\boldsymbol{\Sigma}_i$ for a given multivariate (combinatorial) state i is computed as follows: From the combination of univariate state calls (U or M) of all samples, we pick those bins that show combinatorial state i. The read counts $\boldsymbol{x}_t$ in those bins are transformed to $\boldsymbol{z}_{it}$ using equation 4.9 and the correlation matrix $\boldsymbol{\Sigma}_i$ is calculated from the transformed read counts.

Similarly to the univariate Hidden Markov Model, we use the Baum-Welch algorithm to obtain the best fit for the transition probabilities and posterior probabilities of being in a given state. However, the emission densities remain fixed in the multivariate case. We assign a combinatorial state to each bin by maximizing over the posterior probabilities. We found it useful to transform posterior probabilities for each combinatorial state ("posteriors-per-state") into posterior probabilities of peak calls for each experiment ("posteriors-per-mark"), such that a cutoff can be applied instead of maximizing over the posteriors. We found that our algorithm had a very high sensitivity for detecting peaks when maximizing over the "posteriors-per-state". To increase specificity, a strict cutoff (*e.g.* 0.9999) can be applied on the "posteriors-per-mark".

Sliding windows

The first step of our approach is the binning of the genome into non-overlapping windows. Obviously, a very small bin size will yield high-resolution peak calls, but this comes at the cost of a decreased signal-to-noise ratio. A possible solution that keeps the signal-to-noise ratio constant while allowing more fine grained peak calls are sliding windows with bin size B and step size S. For instance, we can define a bin size of $B = 1000$ bp and a step size of $S = 200$ bp. This induces $B/S = 5$ different binnings with bin size $B = 1000$ bp, with offsets of 0 bp, 200 bp, 400 bp, 600 bp and 800 bp. We estimate parameters and posteriors for the Hidden Markov Model with offset 0 bp, and re-estimate only the posteriors for the binnings with offset. Finally, for each bin in an $S = 200$ bp binning, we assign the state with the highest posterior probability over the different offsets.

Inclusion of replicates

The `chromstaR` formalism offers an elegant way to include replicates. For a single ChIP-seq experiment, there are two states - unmodified (background) and modified (peaks). For an arbitrary number of N experiments, there are thus 2^N combinatorial states. The same is true for an arbitrary number of replicates R, which would yield 2^R combinatorial states. However, in the case of replicates, the number of states can be fixed to 2, such that all replicates are forced to have the same state (either peak or background). Treating replicates in this way allows to find the most likely state for each position considering information from all replicates without prior merging.

Integration of chromatin input experiments

Chromatin input experiments serve as controls for bias in chromatin fragmentation and variations in sequencing efficiency [80]. We optionally integrate this information by modifying the vector of read counts that serves as the observable in the Hidden Markov Model. Let $\boldsymbol{x}$ be the vector of read counts along the genome for the ChIP-seq experiment, and $\boldsymbol{y}$ be the vector of read counts for the input experiment. Let furthermore $\boldsymbol{y}_P$ be the vector $\boldsymbol{y}$ without zero read counts. In a first step, we null regions with artificially high read counts, *e.g.* repetitive regions around centromers, by setting $x_t = 0$ for all bins t where $y_t >= c$. c is defined as the 99.99% quantile of $\boldsymbol{y}_P$. In a second step, we calculate a corrected read count $\boldsymbol{x}$' as

$$\boldsymbol{x}' = \boldsymbol{x} \cdot min\left(\frac{mean(\boldsymbol{y}_P)}{runmean(\boldsymbol{y})}, \ 1.5 \right) \ , \qquad \text{(eq. 4.10)}$$

where runmean( ) calculates a running mean of 15 bins. This operation modifies the read count $\boldsymbol{x}$ in such a way that $\boldsymbol{x}$' is decreased in bins which have more than average counts in the input and increased in bins that have less than average counts in the input.

Univariate approximation of multivariate state distribution

`chromstaR` offers the possibility to restrict the number of combinatorial states to any number lower than 2^N, where N is the number of ChIP-seq experiments. Because the first step of the `chromstaR` workflow is a univariate peak calling, we can combine those peak calls ad-hoc into combinatorial states and use their ranking to determine which states to use for the multivariate peak calling. Most systems seem to be sparse in their combinatorial patterns, *i.e.* do not utilize the full combinatorial state space [57], therefore, it is often not necessary to run the multivariate part with all 2^N combinations. For instance, for the human Hippocampus tissue with seven marks (see Data Acquisition, page 81), running the multivariate with only 30 instead of 128 states recovers 98.2% of correct state assignments compared to the full 128 state model, and choosing 60 instead of 128 states recovers already 99.5% of correct state assignments compared to the full 128 state model (Figure 4-6). A computational feasible number of 128 combinatorial states will be sufficient for most applications.

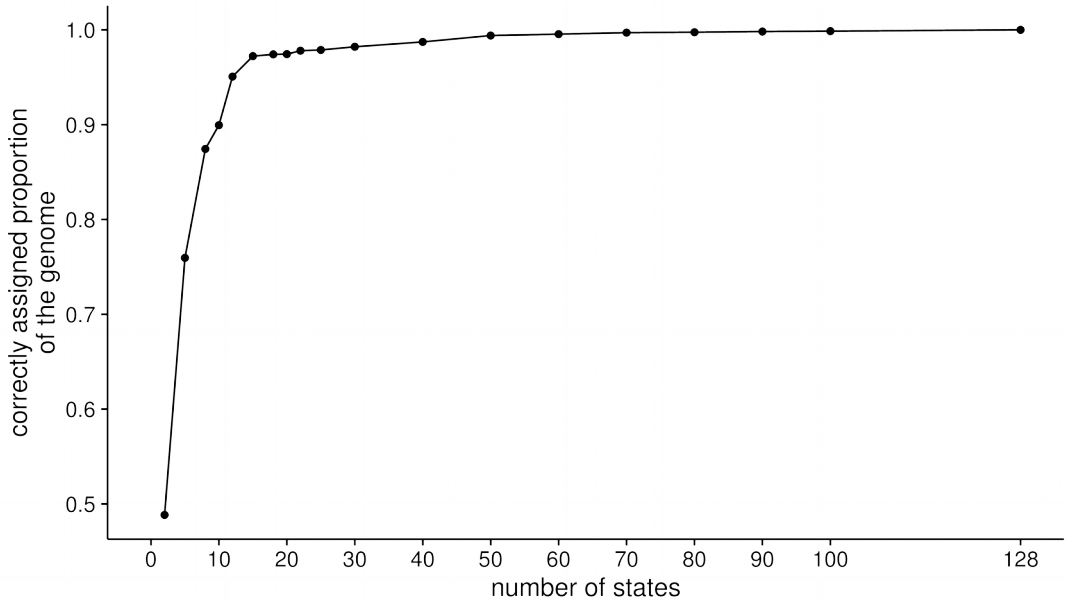


Figure 4-6 | Approximation of multivariate state distribution with less than 2^N states. For the Hippocampus data with seven marks there are $2^7 = 128$ possible combinatorial states (last data point). The figure shows the proportion of the genome that is correctly assigned compared to the full 128-state model (y-axis) if the multivariate is run with fewer than 128 states (x-axis). (Source: Taudt et al. 2016, [51])

Results

Validating univariate peak calls

The univariate part of our model (see page 59), which serves as the basis for the construction of the multivariate model, provides high-quality peak calls that measure up against existing methods. We compared our method with other commonly used peak callers, `MACS2` [81], `Sicer` [82], `BCP` [83] and `Music` [84] using publicly available datasets of qPCR validated regions [85]. Performance of all methods was compared on two datasets, one for H3K4me3 (narrow profile, no input, GSM307618), and one for H3K27me3 (broad profile, with input, GSM721294). The H3K4me3 dataset had 33 qPCR validated regions and the H3K27me3 dataset had 197 qPCR validated regions. The ChIP-seq datasets were analyzed with the standard or recommended settings of each peak caller for narrow and broad marks, respectively. Each base pair was assigned a score by the algorithm and this output was used to compute receiver operator characteristic (ROC) curves and area-under-curve (AUC) values [86]. The performance of `chromstaR` for these datasets in terms of the AUC is equal or better than that of the other methods (Figure 4-7).

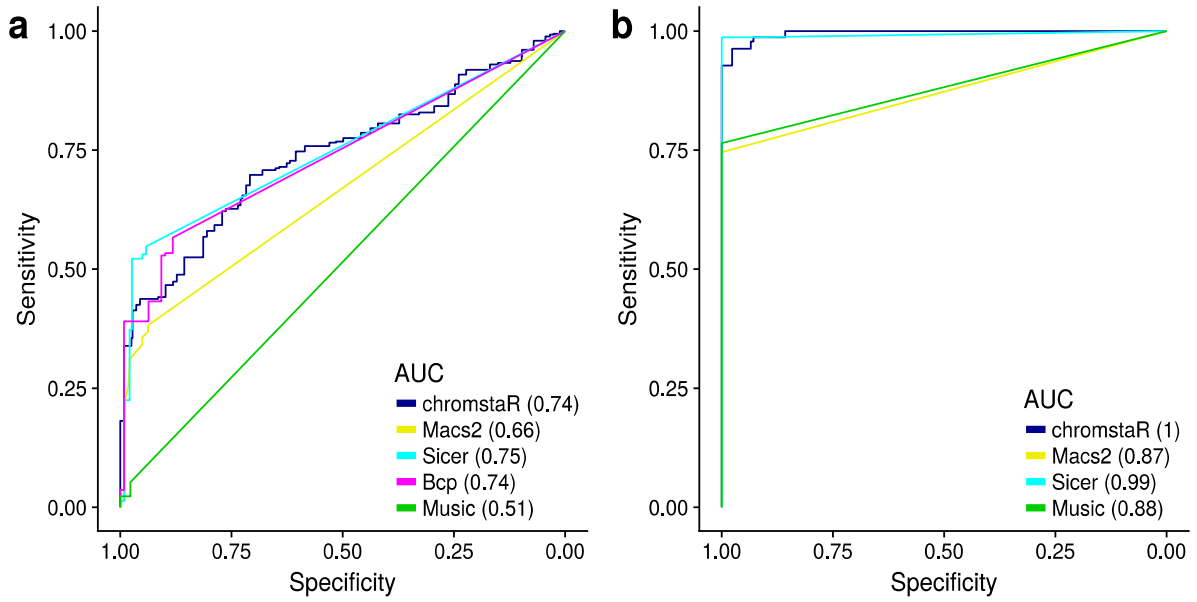


Figure 4-7 | Validation of univariate peak calls. Receiver operator characteristics curves show the sensitivity and specificity of different peak calling methods evaluated using qPCR validated regions from [85] for **a** | H3K27me3 with a broad profile and **b** | H3K4me3 with a narrow profile. Area-under-curve (AUC) values are shown in the legend. (Source: Taudt et al. 2016, [51])

Application 1: Mapping combinatorial chromatin states in a reference tissue

Lara-Astiaso et al. [52] measured four histone modifications (H3K4me1, H3K4me2, H3K4me3 and H3K27ac) and gene expression in 16 mouse hematopoietic cell lines and their progenitors (Figure 4-4). All four marks have a relatively narrow ChIP-seq profile. The authors' goal was to document the dynamic enhancer landscape during hematopoietic differentiation. With four measured histone modifications there are $2^4 = 16$ possible combinatorial states defined by the presence/absence of each of the modifications. In order to provide a snapshot of the genome-wide distribution of these combinatorial states in a given cell-type, we applied `chromstaR` to the ChIP-seq samples collected from monocytes (see SI-Figure 4-20 for the analysis of other cell types). In the following we introduce a shorthand notation where combinatorial states are denoted between brackets [ ] and each mark is abbreviated by its chemical modification. For example, the combination [H3K4me1+H3K4me2+H3K27ac] will be abbreviated as [me1/2+ac]. If we use the full name of a mark (*e.g.* "H3K4me1") we are referring to the mark in a classical, non-combinatorial, context. See Figure 4-8|d for all combinations with shorthands.

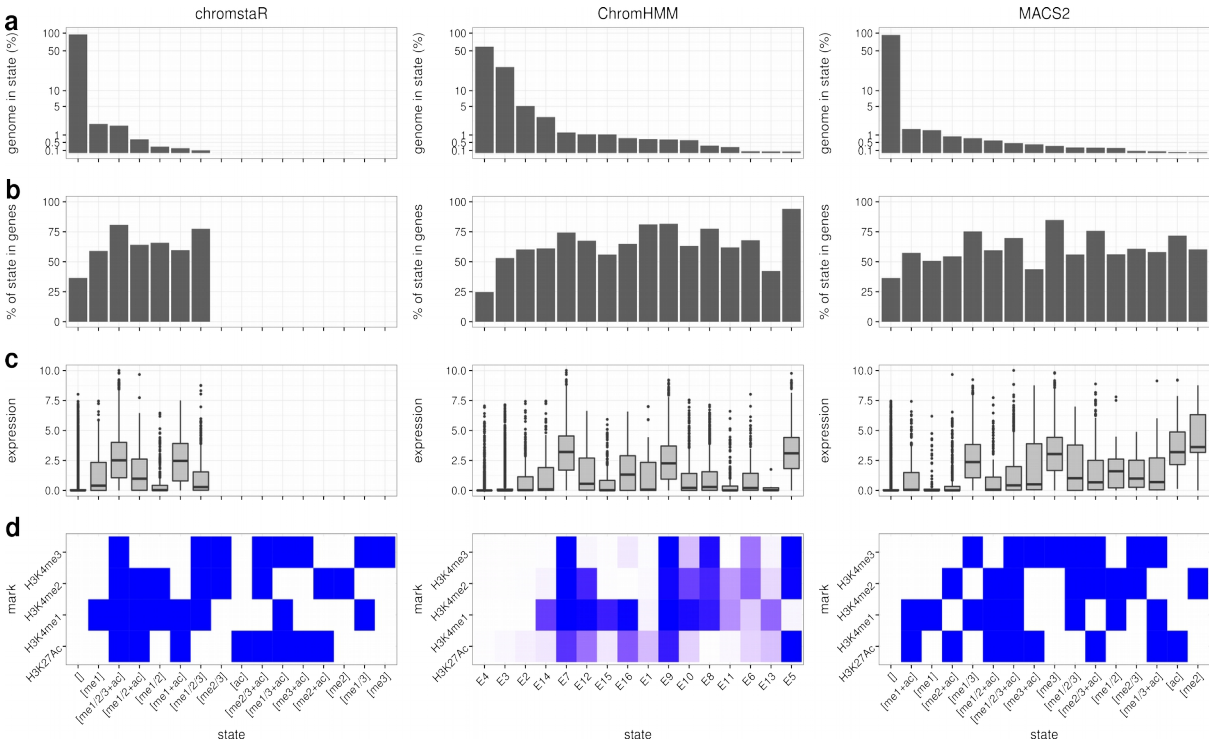


Figure 4-8 | Chromatin states in monocytes. **a** | Genomic frequency, *i.e.* the percentage of the genome that is covered by the chromatin state. The sum over all states equals 100%. **b** | Overlap with known genes. **c** | Expression levels of genes whose transcription start site (TSS) overlaps the chromatin state. **d** | Heatmap showing the chromatin state definition. Histones in `chromstaR` and `MACS2` states are either present (blue) or absent (white). `ChromHMM` states have a continuous emission probability from zero (white) to one (blue). (Source: Taudt et al. 2016, [51])

`chromstaR` found that many of the 16 possible combinatorial states were nearly absent at the genome-wide scale, with 7 of the 16 states accounting for nearly 100% (> 99.99%) of the genome (Figure 4-8|a). This observation indicates that the "histone code" defined by these four histone modifications is much less complex than theoretically possible, perhaps as a result of biochemical constraints on the co-occurance of certain modifications on the same or neighboring aminoacid residues. However, some of the discovered chromatin states display “incompatible” combinations (the ones displaying more than two modifications on the same histone and residue, such as for example [me1/2/3]). Re-analysis of the data with smaller bin sizes finds eight of the 16 states present in the genome, with a smaller frequency of incompatible states (Figure 4-9). These results show that these states are in part due to having pooled data from several nucleosomes into the same bin, but are probably also caused by antibody cross-reactivity and residual cell heterogeneity.

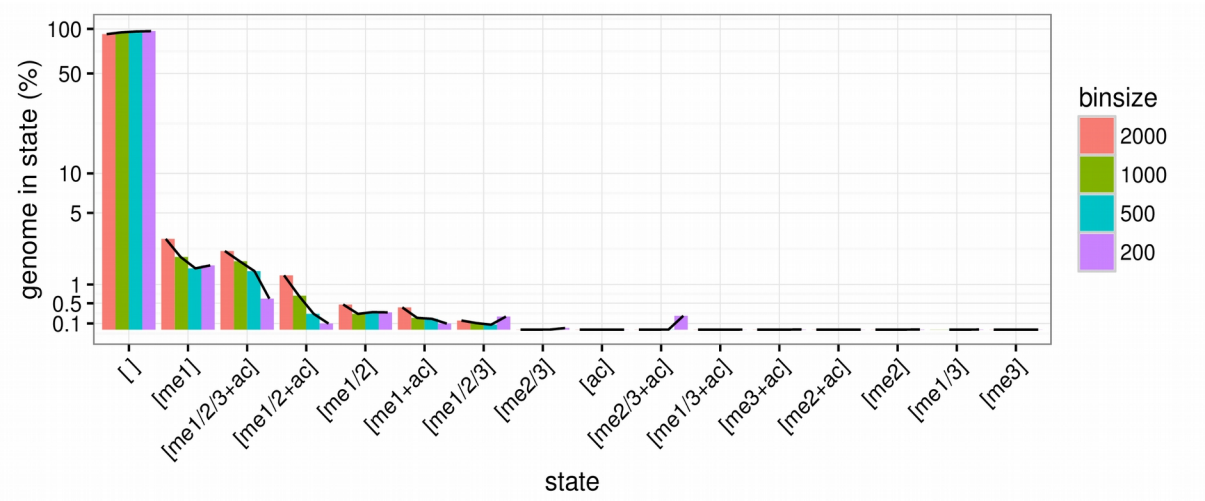


Figure 4-9 | Effect of binsize on chromatin state frequencies in monocytes. With decreasing binsize and higher resolution, the frequency of all non-empty states decreases, while the frequency of the empty state [ ] increases. This effect is mainly due to the decrease in the frequency of acetylation, and also to the decrease in the frequency of mechanistically incompatible states. Multiple methylation states still occur at binsize 200 and are probably caused by cross-reactivity of the antibody and cell heterogeneity. (Source: Taudt et al. 2016, [51])

The empty state [ ], which we here define as the simultaneous absence of all measured marks at a given genomic position, was the most frequent state, covering 94.8% of the genome. The high prevalence of this state reflects the fact that Lara-Astiaso et al. [52] focused on marks with a narrow profile that had previously been shown to occur proximal to genic sequences [59], [60], [87]. Indeed, only 36% of the empty state overlapped known genes while the remaining 64% mapped to non-genic regions throuhgout the genome, and probably tag other (unmeasured) histone modifications, such as repressive heterochromatin-associated marks.

In order to evaluate chromatin state frequencies on a data set with a mixture of broad and narrow histone modifications, we analyzed human Hippocampus tissue data from the Epigenomics Roadmap with seven marks [27] and IMR90 cell line data from the ENCODE project with 26 marks [66] (see Data Acquisition, page 81 and Multivariate peak-calling, page 81). In the Hippocampus data we found that only 21 out of the 128 possible combinatorial states were necessary to explain more than 99% of the epigenome, and indeed the empty state covered only 32% of the genome (SI-Figure 4-21). Moreover, in the lung fibroblast cell line we found that from the possible 67 million states only 0.02% (~12000) are needed to explain more than 95% of the genome, while the empty state covered only 16% of it, showing that when more marks are included the percentage of the genome in the empty state decreases [57].

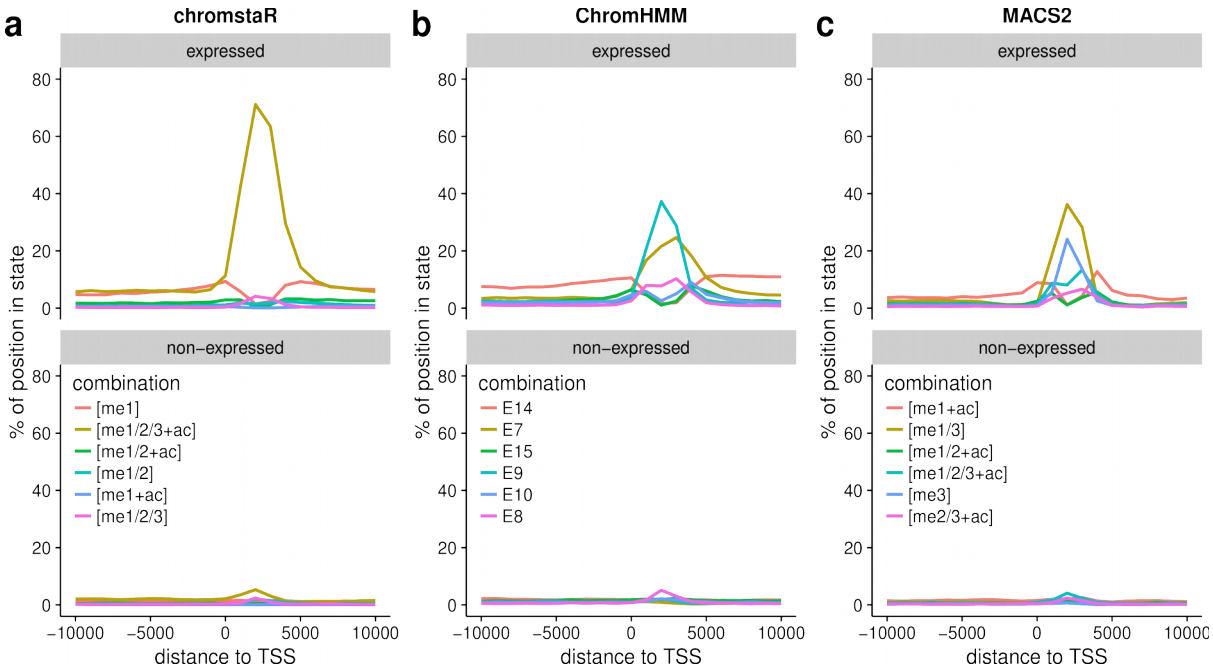


Figure 4-10 | Fold enrichment of chromatin states around transcription start sites (TSS) of expressed (top) and non-expressed (bottom) genes. Shown are the enrichment profiles for the 6 states that are most enriched around TSS. `chromstaR` consistently assigns state [me1/2/3+ac] to expressed TSS and has a higher sensitivity for detecting those, while the other two methods assign two states with lower sensitivity. (Source: Taudt et al. 2016, [51])

Contrary to the empty state, on average 68% (range: 59-81%) of the genomic regions found to be in one of the 6 most frequent (non-empty) combinatorial states in mouse monocytes overlap known genes (Figure 4-8|b), thus suggesting an active role in the regulation of gene expression. To assess this, we examined the combinatorial state profiles of the 6 most frequent states relative to the transcription start site (TSS) of expressed and non-expressed genes (Figure 4-10|a). In contrast to non-expressed genes, expressed genes were clearly characterized by the presence of state [me1/2/3+ac] proximal to the TSS. This is consistent with previous reports that have used H3K4me3 together with H3K27ac to tag active promoters [88]. However, our analysis also uncovered a more subtle enrichment of state [me1] shouldering the TSS (Figure 4-10|a).

We found that 18% of [me1] sites occur in regions directly flanking state [me1/2/3+ac] and 61% of all [me1] can be found within 10 kb of [me1/2/3+ac] sites (see Figure 4-11 for an example). These two states therefore constitute a single, broad chromatin signature that defines a subset of expressed genes. Interestingly, this subset of genes had significantly higher expression levels ($p \approx 10^{-101}$, Wilcoxon rank-sum test) and distinct GO terms compared with genes marked only by the active promoter state (*i.e.* [me1/2/3+ac] at the TSS and no [me1] in flanking regions, Figure 4-12 and Table 4-1). This observation suggests that the co-occurance of [me1/2/3+ac] and [me1] in broad regions surrounding the TSS marks what may be called “enhanced” active promoters ([me1/2/3+ac]+[me1]).

Table 4-1 | The first 10 significant gene ontology terms for TSS overlapping the [me1/2/3+ac] state with the [me1] state flanking it, versus the TSS overlapping the [me1/2/3+ac] state. Numbers indicate the binomial false discovery rate (BinomFdrQ) as reported by GREAT. (Source: Taudt et al. 2016, [51])

	[me1/2/3+ac] + flanking [me1]		[me1/2/3+ac]	
1	posttranscriptional regulation of gene expression	5.75E-25	RNA processing	1.40E-36
2	regulation of translation	1.57E-21	ncRNA metabolic process	6.75E-34
3	peptidyl-lysine modification	1.69E-17	ncRNA processing	2.57E-29
4	microtubule nucleation	2.80E-13	DNA repair	1.14E-25
5	mRNA transport	1.23E-10	ribosome biogenesis	3.36E-17
6	RNA localization	3.61E-10	rRNA metabolic process	9.14E-17
7	RNA transport	7.01E-10	tRNA metabolic process	3.14E-16
8	GPI anchor biosynthetic process	1.01E-09	rRNA processing	1.85E-15
9	negative regulation of translation	1.16E-09	protein folding	3.74E-14
10	DNA replication	1.60E-09	tRNA processing	1.02E-11

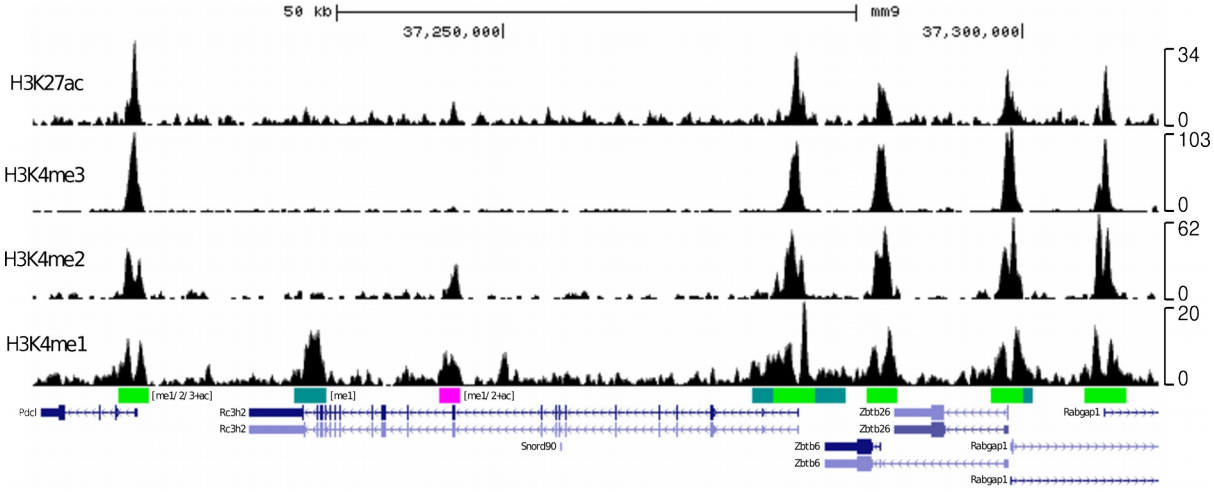


Figure 4-11 | Example of me1-flanked promoter signature showing a 100 kb region on chromosome 2 with several active promoter signatures [me1/2/3+ac] (green) flanked by the [me1] signature (teal). The four black tracks show normalized signal coverage profiles for the four measured histone modifications. The colored track below shows combinatorial state calls, and the bottom track shows genes. Promoter signatures coincide with gene starts. (Source: Taudt et al. 2016, [51])

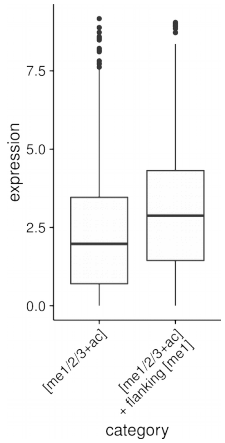


Figure 4-12 | Expression levels of genes whose transcription start site (TSS) shows either the [me1/2/3+ac] signature alone or the [me1/2/3+ac] signature flanked by [me1]. TSS flanked by [me1] show significantly higher expression levels ($p \approx 10^{-101}$, Wilcoxon rank-sum test). (Source: Taudt et al. 2016, [51])

To compare the results obtained with chromstaR to other computational approaches, we first re-analyzed replicate datasets from the mouse monocytes [52], the human Hippocampus [27] and the lung fibroblast cell line data [66] using several publicly available methods [55], [67], [70], [71] (SI text, page 83). chromstaR is the method that provides the best performance in assigning a consistent segmentation between replicates (SI-Figure 4-18|c) and in detecting regions with high read count fold change as differential (SI-Figure 4-18|d). Among the alternative methods, ChromHMM provides the best performance and flexibility [67], we will therefore use it in the following for comparison purposes. ChromHMM employs a multivariate HMM to classify the genome into a preselected number of probabilistic chromatin states, and was used to annotate the epigenome in the ENCODE [66] and Epigenomics Roadmap [27] projects. It therefore offers a method to segment the genome into a set of probabilistic chromatin states that can then *a posteriori* be interpreted at the biological level. We also compare the results obtained with chromstaR to the ones obtained using MACS2 [81], because it is one of the most widely used univariate peak callers.

When using a multivariate segmentation method like ChromHMM, the number of chromatin states needs to be decided beforehand, which is difficult as this number is rarely known *a priori*. In the absence of detailed guidelines we fitted a 16 state model to the mouse hematopoietic data. Our comparison uncovered substantial method-specific differences in state frequencies (Figure 4-8). Both ChromHMM and MACS2 found all 16 states present in the genome with more than 0.01% genome coverage. To understand how state-calls compared between methods, we evaluated to which extent the states detected by one method coincided with those detected by the other method(s) (SI-Figure 4-22). Most notable, we found that genomic regions corresponding to chromstaR's active promoter state [me1/2/3+ac] were assigned to two alternative states (E7 and E9) by ChromHMM. These latter two states were very similar in terms of their emission densities, but significantly different at the level of gene expression ($p \approx 10^{-90}$, t-test, Figure 4-8|c). Moreover, chromstaR's single empty state [ ] corresponded to two functionally similar (nearly) empty states (E3, E4) detected by ChromHMM. A third almost empty state E2 with weak H3K27ac signal had slightly higher expression levels than the other two empty states and also overlapped with chromstaR's empty state [ ] (SI-Figure 4-22|b). These state redundancies highlight the difficulty in selecting the number of chromatin states for ChromHMM, for without extensive manual curation it is difficult to know if two states are truly redundant (likely E3 and E4) or if they are biologically different on some level (E7 and E9).

Although MACS2 is not designed for multivariate analysis, we constructed *ad hoc* combinatorial state calls from the univariate analyses obtained from each ChIP-seq experiment to illustrate the problems of this commonly used analysis technique. As expected, MACS2 results were noisy: many of the combinatorial states detected by chromstaR showed very heterogenous state calls with MACS2 (SI-Figure 4-22|a). For instance, a considerable proportion (35%) of genomic regions detected by chromstaR as being in the active promoter state [me1/2/3+ac] were assigned to another promoter state (containing H3K4me3) by MACS2. We suspect that this is due to the limitations of MACS2 in calling

broader marks (*e.g.* H3K4me1) or moderate enrichment with the default parameters, which results in frequent missed calls for individual modifications, and subsequently also in the limited detection of "complex" combinatorial states such as [me1/2/3+ac] that are defined by the presence of all modifications.

To better understand the functional implications of the state frequency and state pattern differences between these methods, we evalute the chromatin state signatures of both `ChromHMM` and `MACS2` around TSS of expressed and non-expressed genes (Figure 4-10|b,c). In contrast to `chromstaR`, chromatin signatures obtained by the other two methods did not as effectively distinguish these two classes of genes, suggesting that `chromstaR` has a higher sensitivity for detecting these signatures (Table 4-2, see SI page 81 for details about calculation of performance).

Table 4-2 | Performance for detecting expressed TSS. (Source: Taudt et al. 2016, [51])

	Sensitivity	Precision	F1-score
chromstaR	0.71	0.97	0.82
MACS2	0.6	0.98	0.75
ChromHMM	0.59	0.98	0.73

Application 2: Differential analysis of combinatorial chromatin states

In order to understand combinatorial chromatin state signatures that are specific to a given cell type or disease state, it is necessary to compare at least two different tissues with each other, or a case and a control. In this context, the goal is to identify genomic regions showing differential (or non-differential) combinatorial state patterns. Such differential patterns are indicative of regions that underly the tissue differences and are therefore of substantial biological or clinical interest. `chromstaR` solves this problem by considering all 2^{2N} possible combinatorial/differential chromatin states (Figure 4-3|c), where N is the number of histone modifications measured in both conditions. Out of the 2^{2N} states, 2^N are non-differential and 2^{2N}-2^N are differential.

We anlayzed two differentiated mouse hematopoietic cells (monocytes versus CD4 T-cells) from [52], with four histone marks each (H3K4me1, H3K4me2, H3K4me3 and H3K27ac). We found that 5.37% of the genome showed differences in combinatorial state patterns between the two cell types (Figure 4-13|a, example browser shot in Figure 4-14). The most frequent differential regions involved the [me1] combination (2.37%) followed by regions with the [me1/2/3+ac] combination (0.92%). These differences are even more striking when viewed in relative numbers: 59% of the [me1/2/3+ac] sites were concordant between the two cell types, while only 8% of the [me1] sites were concordant. This is in line with previous findings showing that H3K4me1 is highly cell type specific [53], [89]–[91].

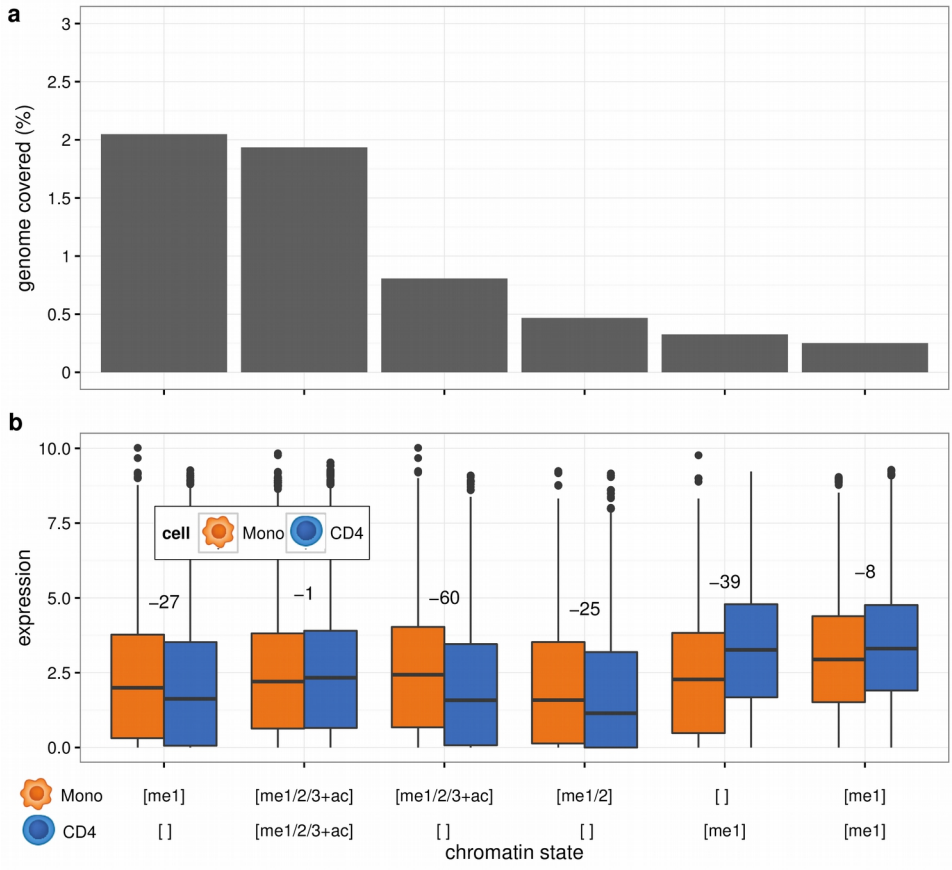


Figure 4-13 | Differential analysis of monocytes and CD4 T-cells. **a** | Genomic frequency of the six most frequent differential chromatin states. Differential chromatin states are shown on the x axes (top: combinatorial state in monocytes, bottom: combinatorial state in CD4). **b** | Expression from monocytes and CD4 cells of genes which overlap the given differential chromatin state. Numbers give the base-10 logarithm of the multiple testing corrected p-value for the expression difference using a Wilcoxon rank-sum test. The more negative the number, the more significant the difference. Genes with differential chromatin signatures (*e.g.* [me1/2/3+ac] in monocytes but not in CD4 T-cells) show highly significant expression differences, whereas similar signatures ([me1/2/3+ac] in both cell types) are not significantly different. (Source: Taudt et al. 2016, [51])

In order to determine if these differences in chromatin play a role in cellular identity, we explored gene expression differences for differential chromatin states. We found that loss of state [me1] as well as of state [me1/2/3+ac] is correlated with a decrease in expression levels (Figure 4-13|b). This is consistent with our previous observation (section Application 1) that [me1/2/3+ac] defines active promoters and [me1] together with [me1/2/3+ac] defines enhanced active promoters (Figure 4-12). To investigate the function of the differential loci, we performed a GO term enrichment of these regions [92] and found an impressive confirmation of cell type identity in the GO terms (SI-Table 4-3): While regions that are marked by [me1/2/3+ac] or [me1] in both cell types show enrichment for general immune cell differentiation terms, regions that are marked with [me1] or [me1/2/3+ac] only in CD4 T-cells show terms such as “T-cell activation and differentiation”. Vice versa, regions that are marked with those signatures in monocytes but not in T-cells show enrichment of terms such as “response to other organism” and “inflammatory response”.

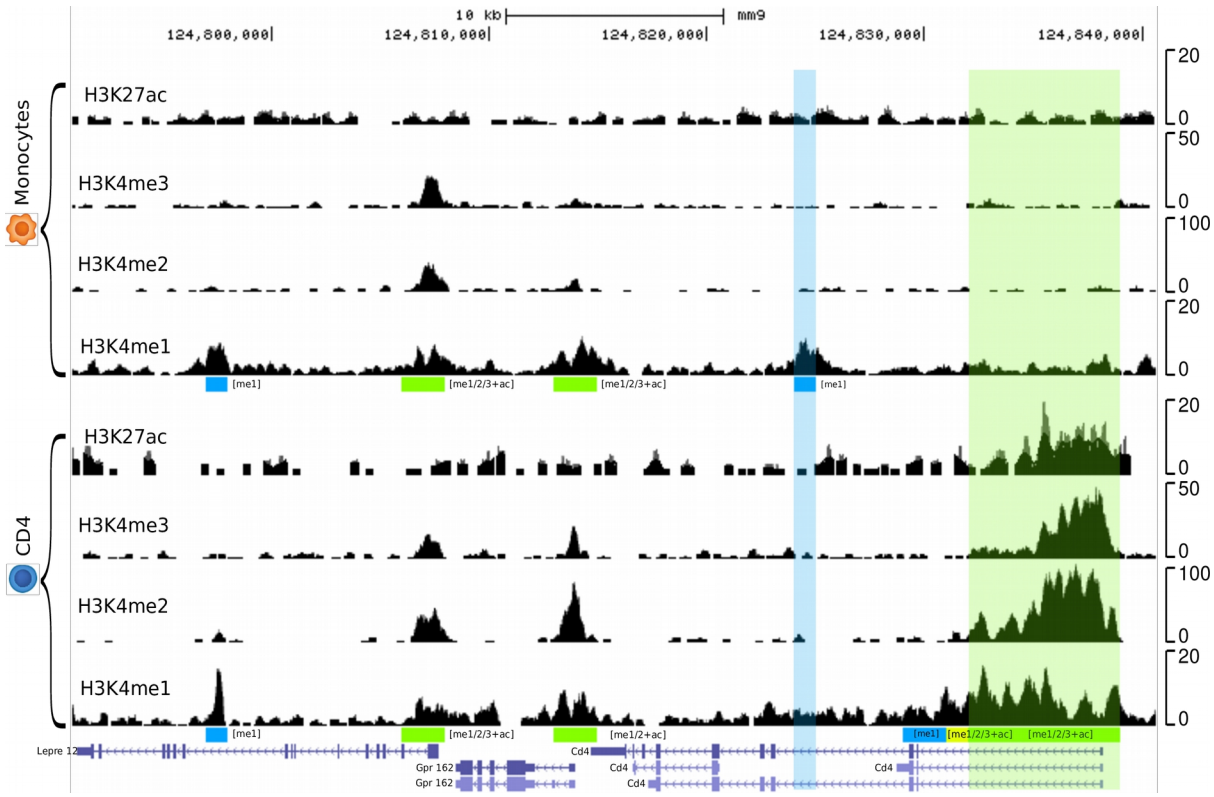


Figure 4-14 | Differential chromatin signature at the Cd4 locus. Example of a differential promoter and enhancer signature at the Cd4 gene. The differential promoter signature [me1/2/3+ac] is only present in CD4 T-cells (shaded green), while the differential enhancer [me1] is present only in monocytes (shaded blue). (Source: Taudt et al. 2016, [51])

Again, we compared our results on the same dataset with `MACS2` [81] and `ChromHMM` [67]. Neither method was specifically designed to deal with differences between combinatorial states, but both tools represent approaches that could have been chosen for that task in the absence of other suitable methods. For both methods, the percentage of the epigenome that was differentially modified was found to be 2.5 times higher than predicted by `chromstaR`, 13.02% for `MACS2` and 13.59% for `ChromHMM`. `MACS2` found most differences (3.90%) in state [me1], followed by the combination [me2+ac] (2.11%). None of these states yielded any significant enrichment in GO terms or showed correlation with expression data (SI-Figure 4-23|c and SI-Table 4-4). The third most frequent differential state was [me1+ac] (1.88%) and this state yielded GO term enrichments which reflect cellular identity. `ChromHMM` predicted two "enhancer-like states" E8 and E9 (SI-Figure 4-23|b) as most differential between cell types (2.71% and 2.54%) which also showed cell type specific terms in the GO analysis (SI-Table 4-5). However, expression analysis showed that `ChromHMM`'s most frequent differential state (CD4:E12 and Mono:E14) corresponded to proximal genes that were transcriptionally nearly inactive (SI-Figure 4-23|b), which raises the question if these differential chromatin states produce cell-specific functional differences.

Application 3: Tracking combinatorial chromatin state dynamics in time

Arguably the most challenging experimental setup is when several histone modifications have been collected for a large number of conditions, such as different cell types along a differentiation tree or different terminally differentiated tissues (Figure 4-4). We consider M conditions with N histone modifications measured in each of them. This leads to 2^N possible combinatorial states per condition, or alternatively to 2^M differential states per mark across all samples. Therefore, the number of possible dynamic combinatorial chromatin states is $2^{M \times N}$. For $M \times N \leq C$ the whole dynamic/combinatorial chromatin landscape is treatable computationally, while for $M \times N > C$ the problem becomes intractable with current computational resources. The value of C is dependent on computational resources, genome length and bin size (see section Limitations, page 76).

We considered again the mouse hematopoietic data from [52], with four histone modifications (H3K4me1, H3K4me2, H3K4me3 and H3K27ac) measured in 16 different cell types during hematopoietic differentiation (stem cells, progenitor and terminally differentiated cells). We explored the chromatin dynamics during the differentiation process for every hematopoietic branch (Figure 4-4|a): first, long term hematopoietic stem cells (LT-HSC) are transformed into short term hematopoietic stem cells (ST-HSC) and further into multipotent progenitors (MPP). The MPP cells differentiate into the several common lineage oligopotent progenitors, giving rise to the three different hematopoietic branches (myeloid, leukocyte and erythrocyte). Finally, after another one or two stages, cells become fully differentiated at the bottom of the tree. Every branch from root to leaf consists therefore of four histone marks in five or six time points, with $2^{M \times N} = 1048576$ or 16777216 possible dynamic combinatorial chromatin states, respectively. Because this number is computationally intractable, we implemented the following two-step approach for each branch: (1) for each of the four histone marks separately, we performed a multivariate differential analysis along the five or six cells in the branch, therefore assigning every bin in the genome to one of the 32 or 64 possible differential states; (2) We reconstructed the full combinatorial chromatin state dynamics by *ad hoc* combining the differential calls of all four marks in step 1, bin by bin (SI-Figure 4-19|a).

Using this two-step approach, we studied the dynamics of the inferred chromatin states over developmental time. We observed an initial increase in the frequency of the [me1] state from the LT-HSC to intermediate progenitor stages, followed by a decrease to the fully differentiated stages (SI-Figure 4-24). This decrease in [me1] was especially pronounced in the lymphoid and erythroid lineage. In the [me1/2/3+ac] signature we found a small but continuous decrease from LT-HSC to terminally differentiated stages. These observations are consistent with the view that chromatin transitions from an open configuration in multipotent cells to a closed configuration in differentiated cells. Figure 4-15 shows two examples of pluripotency genes, Gata2 and Cebpa, that lose their open chromatin configuration in differentiated CD4 T-cells.

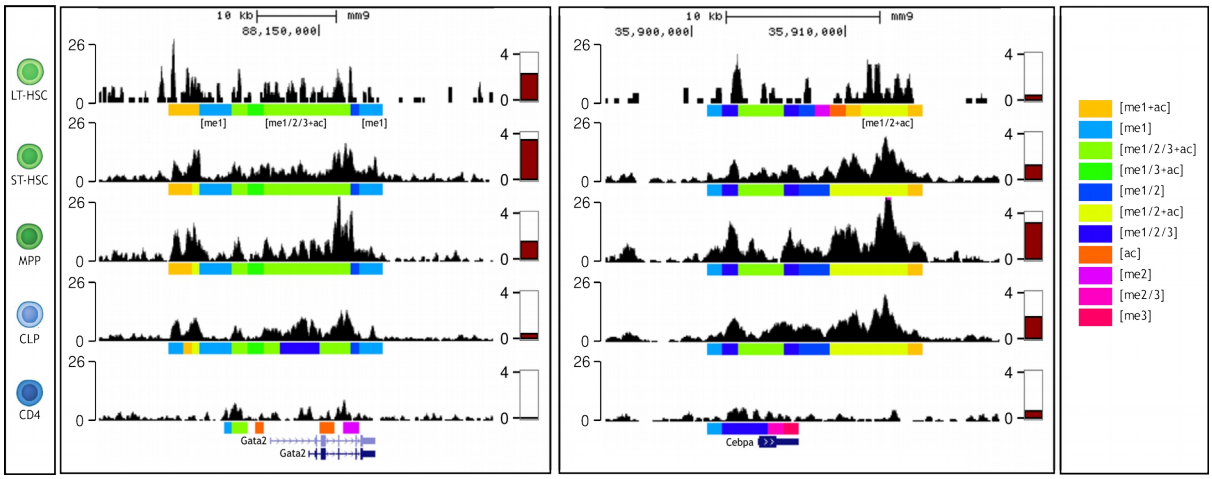


Figure 4-15 | Chromatin state transitions at **a** | Gata2 and **b** | Cebpa. Black genome browser tracks show H3K4me1 levels and combinatorial chromatin states as determined by chromstaR below. Red bars on the right of each track indicate normalized expression levels of Gata2 and Cebpa, respectively. Both Gata2 and Cebpa are involved in maintenance of pluripotency in stem cells and transition from an open into a closed chromatin configuration during differentiation. (Source: Taudt et al. 2016, [51])

We next explored the specific dynamic chromatin state transitions that occur in every region of the genome during the differentiation process. We found that the majority of all possible dynamic chromatin state transitions were not present in this system. For example, in the CD4 T-cell branch of the hematopoietic tree there are 5 developmental time points and at each stage 16 combinatorial states can be theoretically present. This leads to $16^5 = 1048576$ potential transitions between combinatorial states in this branch. However, we found only 1086 different chromatin transitions and the first most frequent 99 transitions (with frequency $\geq 0.01\%$) already involved 99.60% of the genome. To summarize these transitions further, we grouped them into 4 different classes: (1) “Empty” transitions, *i.e.* those regions that have no histone modification in any of the developmental stages. (2) “Constant” transitions, *i.e.* those regions that show the same (non-empty) combinatorial state in all stages of differentiation. (3) “Stage-specific” transitions, *i.e.* those regions that show a combinatorial state only in a subset of differentiation stages and are in the “empty” state otherwise. (4) All other transitions (see Figure 4-16 for examples). In the CD4 T-cell branch, 85.98% of the genome has no measured chromatin signature in all 5 stages (class~1). The constant transitions (class~2) comprise 5.87% of the genome, stage-specific transitions 5.69% (class~3) and all other transitions 2.46% (class~4). Altogether, only 8.15% of the genome changes its chromatin state during differentiation and more than half of these changes are due to changes in the [me1] signature. This signature is highly cell type specific and gains and losses correspond to stage-specific terms in a GO analysis (SI-Table 4-6) and to changes in gene expression (SI-Figure 4-25|a). Among the constant transitions, regions with signature [me1/2/3+ac] mark constitutively expressed genes (SI-Figure 4-25|a). Therefore we expect those regions to be enriched with housekeeping functions, which is confirmed by the GO analysis (SI-Table 4-7).

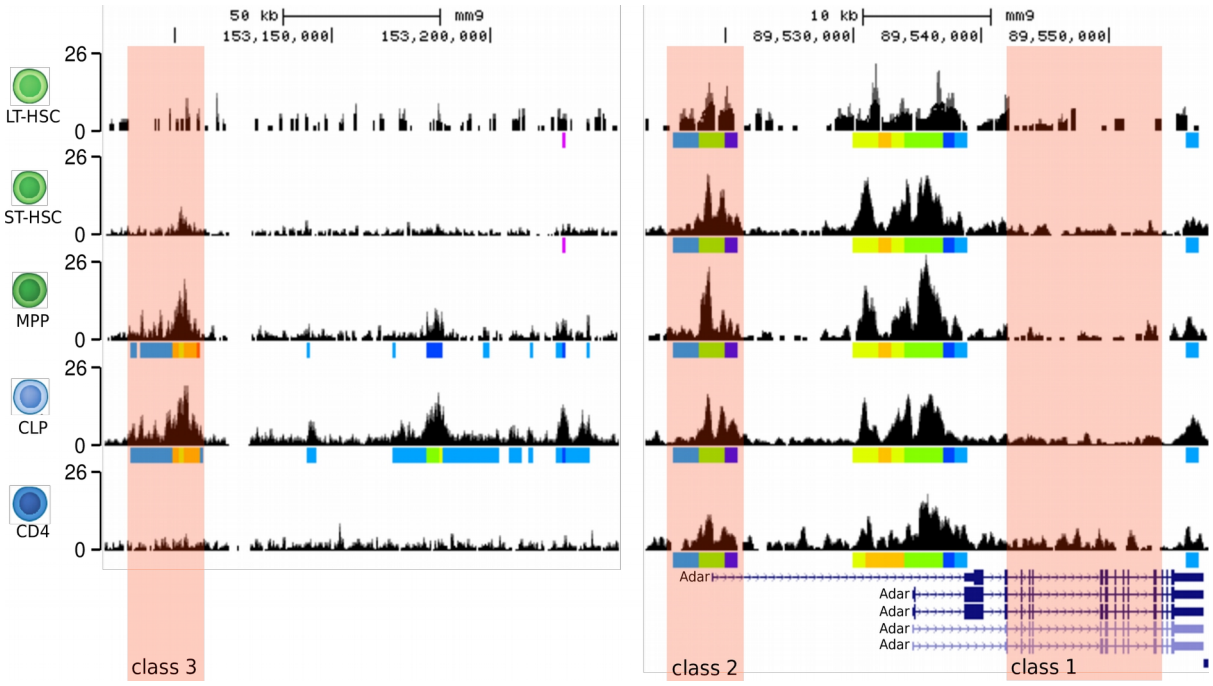


Figure 4-16 | Examples of different classes of transitions (shaded in red): "Empty" (class 1), "constant" (class 2) and "stage-specific" transitions (class 3). Only the H3K4me1 tracks are shown. Combinatorial chromatin states as obtained by `chromstaR` are shown below the H3K4me1 tracks as colored bars. (Source: Taudt et al. 2016, [51])

We compared our results on the CD4 T-cell branch with `MACS2` [81] and `ChromHMM` [67]. Strikingly, `MACS2` found 34470 different chromatin state transitions, with the most frequent 330 (with frequency $\geq 0.01\%$) covering 94.47% of the genome. This large number is expected since `MACS2` is a univariate peak caller and not designed for differential analysis. Furthermore, this dataset represents a differential analysis not between two cell types, but between five different cell types and thus boundary effects (false positives, *e.g.* falsely detected differences) are extremely likely. This interpretation is supported by the expression data, which could not find clear expression differences for the most frequent differentially modified regions (SI-Figure 4-23|c). Also the GO analysis could not identify any significant GO terms. `ChromHMM` found 38288 different state transitions of which the first 656 cover 91.21% of the genome. This large number of transitions is dependent on the number of states that are used to train `ChromHMM`, since extra states will artificially inflate the number of chromatin state transitions. However, consistent with the `chromstaR` predictions, `ChromHMM` predicts many stage-specific enhancer (state E15 and E16) and constant promoter (state E9) regions among the most frequent transitions. The expression profiles associated with those transitions show the expected behaviour (SI-Figure 4-25|b).

Limitations

The number of possible combinatorial states for *N* ChIP-seq experiments is 2^N, meaning that for each additional ChIP-seq experiment the number of combinatorial states doubles. Thus it soon becomes computationally prohibitive to consider all combinatorial states. We found that with current computational resources (Intel Xeon E5 2680v3, 24 cores @ 2.5 GHz, 128GB memory) a practical limit seems to be 256 states (= 8 experiments) with a run-time of several days for a mouse genome (≈ 2.6 10^9 bp) and a bin size of 1000 bp (≈ 2.6M datapoints). We investigated several possibilities to extend the usability of `chromstaR` beyond this limit:

(1) Calculations can be performed for each chromosome separately, and `chromstaR` features an option to perform this calculation in parallel.

(2) For the case of one cell type or tissue where the number of measured histone modifications *N* exceeds the upper limit, `chromstaR` provides a strategy to artificially restrict the number of combinatorial states to any number lower than 2^N. This strategy can yield proper results if the correct states are included, since our results have shown that the majority of combinatorial states are absent in the genome. In order to identify the states which are the most present in the genome, `chromstaR` ranks the combinatorial states based on their presence according to the combination of univariate results from the first step of the `chromstaR` pipeline. This ranking is a good approximation of the true multivariate state-distribution (Figure 4-6).

(3) If there are multiple marks *N* in multiple tissues *M*, and $2^{N \times M}$ is bigger than the maximum number of states that the algorithm can handle computationally, two strategies are possible: One can either perform a differential analysis for each mark and then reconstruct combinatorial states in a classical way (SI-Figure 4-19|a) or one can perform a multivariate peak-calling of combinatorial states for each tissue and then obtain the differences by a simple comparison between tissues (SI-Figure 4-19|b). Both strategies give a different perspective on the data: The former accurately identifies differences between marks, while the combinatorial states might be subject to boundary effects. The latter gives an accurate picture of the combinatorial chromatin landscape, while differences between cells might be overestimated.

(4) The run-time of our algorithm scales linearly with the number of data points, and thus a strategy is to decrease the resolution, *e.g.* to halving the run-time by doubling the bin size.

Discussion

Understanding how various histone modifications interact to determine cis-regulatory gene expression states is a fundamental problem in chromatin biology. It is becoming increasingly clear that certain combinatorial patterns of these modifications define discrete chromatin states along the genome. These chromatin states “encode” cell-specific transcriptional programs, and constitute functional units that are subject to dynamic changes in response to

developmental and environmental cues. Many experimental studies have recognized this and collected ChIP-seq data for a number of histone modifications on the same or different tissue(s) as well as for several developmental time points. Integrative analyses of such datasets often present formidable bioinformatic challenges. Only a few computational methods exist that can analyze multiple histone marks simultaneously in one sample and cluster them into a finite number of chromatin states [46], [55], [67]–[71]. Interestingly, these methods often demand that the user specifies the number of chromatin states beforehand. We find this problematic because this number is often a desired output of the analysis rather than an input. Indeed, the true number of distinct chromatin states in the genomes of various species is subject to debate. In *D. melanogaster* nine chromatin states have been reported [40], while in *A. thaliana* four main states were found [44]. In human, Ernst et al. found 51 states in human T-cells [28]. The Roadmap Consortium reported 15 to 18 states [27]. It remains unclear whether these differences reflect species divergence at the level of chromatin organization, or whether they are due to differences in the assessed chromatin marks and bioinformatic treatment of the data. Without a formal computational framework for defining chromatin states these two possibilities cannot be confidently distinguished.

While multivariate methods such as `ChromHMM` provide possible computational solutions to such questions, these methods employ probablistic chromatin state definitions that are not always readily interpretable. A probalistic interpretation means that different combinatorial histone modification patterns can be simultaneously part of different underlying chromatin states. However, it is not immediately obvious whether the underlying chromatin state are biologically distinct or if they are only statistical entities that are otherwise biologically redundant. Identifying such redundancies is not easy, because of a lack of rules to decide whether two or more chromatin states can or cannot be considered to be equivalent. Such decisions require extensive manual curation of the output, and often presuppose the kind of biological knowledge that one wishes to obtain from the data in the first place.

In contrast to this probabilistic state definition, `chromstaR` outputs discrete chromatin states that are defined on the basis of the presence/absence of various histone modifications. That is, with N histone modifications, it infers all 2^N combinatorial chromatin states (Figure 4-3|a). This interpretation makes it easy to relate the inferred chromatin states back to the underlying histone modification patterns and thus fashions a direct mechanistic link between chromatin structure and function. Moreover, `chromstaR`'s discrete state definition also provides an unbiased picture of the genome-wide frequency of various chromatin states and allows for easy genome-wide summary statistics. For instance, in our analysis of four histone modifications in mouse embryonic stem cells we found that only 7 of the 16 possible states covered almost 100% of the genome, and for the human Hippocampus with seven modifications only 21 of the 128 possible combinatorial states already covered 99% of the genome. Even more extreme, when analyzing 26 marks in a lung fibroblast cell line, we found that only 0.02% of all possible combinatorial states explain 95% of the genome. This striking sparsity in the combinatorial code is interesting and points at certain biochemical

contraints that determine which histone modifications can or cannot co-occur at a genomic locus. Clearly, the genome-wide frequency of inferred combinatorial chromatin states depends on the number and the type of different histone modifications that are used in the analysis. Future studies should systematically investigate the dependency of the number of chromatin states on factors such as number and type of measured histone marks, resolution, organism etc.

By treating discrete combinatorial chromatin states as units of analysis `chromstaR` can also easily track chromatin state dynamics across cell types or developmental time points. In that respect `chromstaR` is unique as no other methdods exist to date that can peform a similar task. To illustrate this we have analyzed four different histone modification in 5 different cell types that are part of the mouse T-cell differentiation pathway. Of the 1048576 combinatorial state transitions, we find that only 99 comprise over 99.60% of the genome. Again, the sparsity in state transitions shows that a few key transitions define the developmental trajectory of T-cell differentiation. One notable transition is the gain or loss of state [me1] near promoters. We note that this state means that only H3K4me1 is present at a locus and no other marks. This is not the same as tracking H3K4me1 modification by itself as this latter mark can appear in a number of different, and often funtionally distinct, chromatin states such as [me1+ac], [me1/2+ac], [me1/2/3]. Hence, focusing on H3K4me1 alone would tag other chromatin state changes that may not be fully informative about T-cell differentiation.

Conclusions

`chromstaR` is a computational algorithm that can identify discrete chromatin states from multiple ChIP-seq experiments and detect combinatorial state differences betweeen cell-types and/or developmental time points. By defining chromatin states in terms of the presence and absence of combinatorial histone modification patterns, it provides an intuitive way to understand genome regulation in terms of chromatin composition at a locus. `chromstaR` can be used for the annotation of reference epigenomes as well as for annotation of chromatin state transitions in well-described developmental systems. The algorithm is written in C++ and runs in the popular R computing environment. It therefore combines computational speed with the extensive bioinformatic toolsets available through Bioconductor [93], [94]. `chromstaR` is freely available from Bioconductor and features a collection of downstream analysis functions.

Applications

Mll2 conveys transcription-independent H3K4me3 in oocytes

Courtney W. Hannah, Aaron Taudt, Huang Jiahao, Lenka Gahurova, Andrea Kranz, Simon Andrews, Wendy Dean, Francis A Stewart, Maria Colomé-Tatché, Gavin Kelsey

Contributions: Data analysis with chromstaR. Bioinformatics analysis.

Nature Structural & Molecular Biology; doi: 10.1038/s41594-017-0013-5

Trimethylation of histone 3 lysine 4 (H3K4me3) is classically thought of as a mark of active promoters and yet it occurs at untranscribed domains. Partial redundancy of H3K4 methyltransferases has made it difficult to delineate the mechanisms underlying genomic targeting of H3K4me3. The oocyte provides an attractive *in vivo* system to investigate this, because extensive acquisition of H3K4me3 occurs in a non-dividing cell and ablation of a single H3K4 methyltransferase, Mll2, prevents most H3K4me3. We developed low-input chromatin immunoprecipitation to interrogate promoter-associated histone modifications H3K4me3, H3K27ac and H3K27me3 throughout oogenesis. In non-growing oocytes, H3K4me3 was restricted to transcriptionally active promoters, but as oogenesis progresses, H3K4me3 accumulates in a transcription-independent manner: targeted to broad inter-genic regions, putative enhancers, and transcriptionally silent H3K27me3-marked promoters. Consequently, thousands of bivalent domains are established during oogenesis. Ablation of Mll2 resulted in loss of transcription-independent H3K4me3, with limited effects on transcription-coupled H3K4me3 or gene expression. Deletion of Dnmt3a/b showed that DNA methylation protects regions from acquiring H3K4me3. Our findings show that there are two independent mechanisms of targeting H3K4me3 to genomic elements, with MLL2 recruited to unmethylated CpG-rich regions independently of transcription.

Histone propionylation is a mark of active chromatin

Adam F. Kebede, Anna Nieborak, Lara Zorro Shahidian, Stephanie Le Gras, Florian Richter, Diana Aguilar Gomez, Marijke P. Baltissen, Gergo Meszaros, Helena de Fatima Magliarelli, Aaron Taudt, Raphael Margueron, Maria Colomé-Tatché, Romeo Ricci, Sylvain Daujat, Michiel Vermeulen, Gerhard Mittler, Robert Schneider

Contributions: Data analysis with chromstaR.

Nature Structural & Molecular Biology; doi: 10.1038/nsmb.3490

Histones are highly covalently modified, but the functions of many of these modifications remain unknown. In particular, it is unclear how histone marks are coupled to cellular metabolism and how this coupling affects chromatin architecture. We identified histone H3 Lys14 (H3K14) as a site of propionylation and butyrylation *in vivo* and carried out the first systematic characterization of histone propionylation. We found that H3K14pr and H3K14bu are deposited by histone acetyltransferases, are preferentially enriched at promoters of active genes and are recognized by acylation-state-specific reader proteins. In agreement with these findings, propionyl-CoA was able to stimulate transcription in an *in vitro* transcription system. Notably, genome-wide H3 acylation profiles were redefined following changes to the metabolic state, and deletion of the metabolic enzyme propionyl-CoA carboxylase altered global histone propionylation levels. We propose that histone propionylation, acetylation and butyrylation may act in combination to promote high transcriptional output and to couple cellular metabolism with chromatin structure and function.

Supplemental Material

Data Acquisition

ChIP-seq data for the hematopoietic system (GSE60103) was downloaded from the Gene Expression Omnibus (GEO) and aligned to mouse reference mm9 following the procedure in [52] with bowtie2 (version 2.2.3) [95], keeping only reads that mapped to a unique location. The number of identical reads at each genomic position was restricted to 3. For the expression analysis, we used the provided RNA-seq data (GSE60101). We normalized the read counts by transcript length and scaled them to 1M reads. To reduce the effect of extreme expression values, we applied an arc-sinh transformation on the data.

For the Hippocampus dataset, bed-files with aligned reads were downloaded from ftp://ftp.genboree.org/EpigenomeAtlas/Current-Release/sample-experiment/Brain_Hippocampus_Middle/ for donors number 112 and 149 and seven histone marks H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3, H3K9me3 and H3K9ac.

Bed-files with aligned reads for the IMR90 dataset were downloaded from ftp://ftp.genboree.org/EpigenomeAtlas/Current-Release/sample-experiment/IMR90_Cell_Line for 26 histone marks (H2AK5ac, H2BK120ac, H2BK12ac, H2BK15ac, H2BK20ac, H2BK5ac, H3K14ac, H3K18ac, H3K23ac, H3K27ac, H3K27me3, H3K36me3, H3K4ac, H3K4me1, H3K4me2, H3K4me3, H3K56ac, H3K79me1, H3K79me2, H3K9ac, H3K9me1, H3K9me3, H4K20me1, H4K5ac, H4K8ac, H4K91ac).

Enrichment profiles around TSS

We calculated sensitivity (recall), precision and F1-score for the detection of expressed TSS based on the following assumptions: True positives are expressed TSS which are called into the promoter state ([me1/2/3+ac] for `chromstaR`, E7 and E9 for `ChromHMM`, [me1/3] and [me3] for `MACS2`, see Figure 4-10). False negatives are expressed TSS which are not assigned into the promoter state. True negatives are non-expressed TSS which are not assigned into the promoter state. False positives are non-expressed TSS which are assigned the promoter state. We found that `chromstaR` has a higher sensitivity than the other methods and a lower precision. The F1-score is highest for `chromstaR` (Table 4-2).

Multivariate peak-calling

`chromstaR` was run with a bin size of 1000 bp and convergence threshold of *eps* = 0.01 for both the univariate and multivariate part. Univariate fits were checked manually for proper convergence and rerun with different random initial parameter settings where necessary. For all analysis and comparisons, we excluded replicates SRR1521819, SRR1521851 and SRR1521852 (corresponding to CD8-H3K27ac-Rep1, MF-H3K4me1-Rep1, MF-H3K4me1-Rep2) because we could not obtain a proper fit with our method, regardless of initial parameter settings. Replicates were included in the `chromstaR` analysis as separate

ChIP-seq experiments but forced to yield the same state calls (see section “Inclusion of replicates” in the main text). Input experiments were used where available, that is for the Hippocampus and IMR90 data but not for the hematopoietic system. For all analyses shown in “Application 1: Mapping combinatorial chromatin states in a referene tissue” we applied a cutoff on the posterior probabilities of 0.99 to call peaks and combinatorial states with high specificity. We did not apply a cutoff in the differential analysis (Application 2 and 3), since for the purpose of finding differential regions a cutoff increases boundary effects and is thus adversary to the quality of the detected differential regions.

Likewise, `ChromHMM` was run with a bin size of 1000 bp, 16 states, parallel mode and default parameters otherwise. Signal input files for `ChromHMM` were produced by adding the read counts over replicates. `MACS2` (version 2.1.0.20150731) was run with parameters “-g mm --keep-dup all” and default settings otherwise. Replicates were specified separately and handled by `MACS2` internally. For the comparison with `chromstaR` and `ChromHMM`, `MACS2` calls were transformed into a 1000 bp-bin representation by simply extending each peak into its overlapping bin(s). For the hematopoietic data, `chromstaR` and `ChromHMM` were run on chromosomes 1-19 and X, `MACS2` was run with all scaffolds but only chromosomes 1-19 and X retained for analysis. For the human hippocampus and IMR90 cell line data, methods were run on chromosomes 1-22 and X.

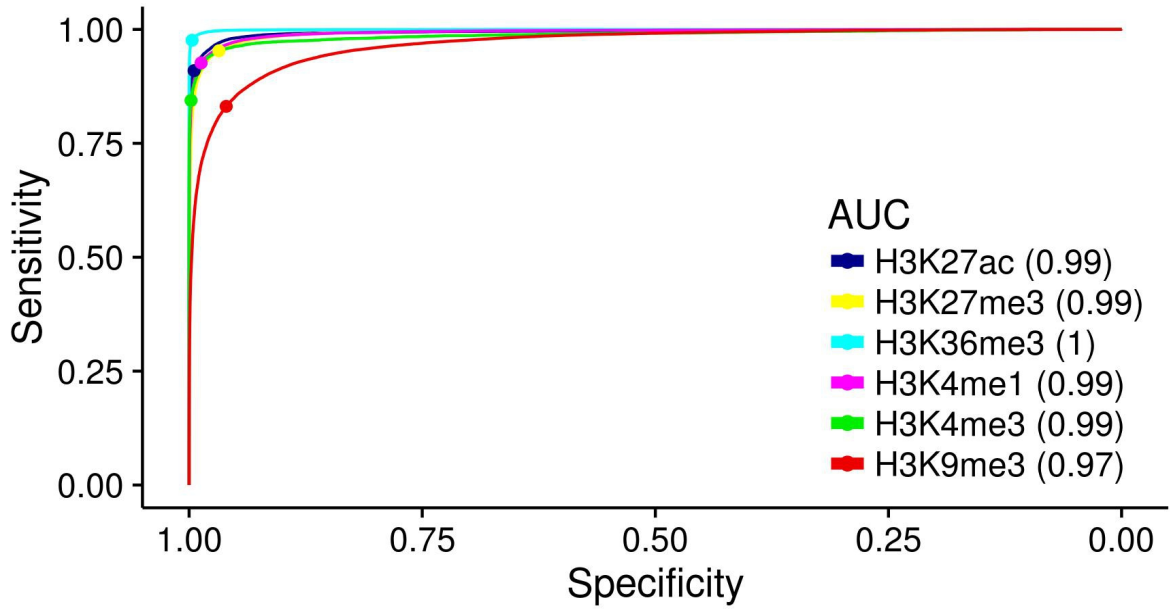
Enrichment analysis and gene ontology

Genomic coordinates were downloaded with `biomaRt` [96], [97] (dataset = mmusculus_gene_ensembl, host = aug2010.archive.ensembl.org) and the first three basepairs of each gene were defined as coordinates for the transcription start site. For the overlap of chromatin states with genes (Figure 4-8|b) we included the promoter region defined as 2 kb upstream of each gene in the gene definition. Gene ontology enrichment was performed with `GREAT` [92] using the whole genome as background set. Only significant terms were retained with the following thresholds: BinomFdrQ < 0.05, HyperFdrQ < 0.1, RegionFoldEnrich >2. Presented terms in all tables are from category “GO Biological Process” and ordered by BinomFdrQ with the most significant results on top.

Simulation study for multivariate peak calls

To assess performance of the multivariate part of our algorithm we performed a simulation study, since no validated experimental datasets exist for that purpose. ChIP-seq reads were simulated for six histone modifications using our univariate Hidden Markov Model with real parameters from the Hippocampus dataset for the transition matrix and emission distributions. Reads were simulated for chromosome 1. We then applied chromstaR on this simulated data and calculated receiver-operater-characteristic (ROC) curves and area-under-curve (AUC) values for the multivariate peak calls. SI-Figure 4-17 shows sensitivity and specificity for peak calls determined by maximizing over the “posteriors-per-state” (dots) and

ROC curves obtained by varying the threshold over the transformed “posteriors-per-mark” (lines). Note that we have not included other methods in this simulation study because a direct comparison of combinatorial and probabilistic chromatin states at the level of peak calls is impossible. Please see the next section for a comparison on real datasets.



SI-Figure 4-17 | Simulation study for multivariate peak calls. Receiver operator characteristic curves show the sensitivity and specificity of `chromstaR`'s multivariate peak-calls for a simulated dataset based on real parameters. Area-under-curve (AUC) values are shown in the legend. Dots indicate values if peaks are called by maximizing over ``posteriors-per-state'' and lines show the values for different cutoffs applied on the transformed ``posteriors-per-mark''. (Source: Taudt et al. 2016, [51])

Comparison with other multivariate methods

We compared chromstaR with 4 other multivariate methods, `ChromHMM` [67], `jMOSAiCS` [55], `Spectacle` [70] and `EpiCSeg` [71]. `chromstaR` and `jMOSAiCS` calculate combinatorial chromatin states, while the other three methods – `ChromHMM`, `Spectacle` and `EpiCSeg` – calculate probabilistic chromatin states.

Since different chromatin state definitions (combinatorial or probabilistic) lead to very different chromatin states, it is not possible to directly compare the results between the utilized methods. For this reason we assess the performance of every method separately in the calling of chromatin state maps in pairs of replicates. We chose three data sets for this comparison:

- The mouse monocyte data [52] used in Application 1 consisting of four histone marks.
- The human hippocampus data from the Epigenomics Roadmap [27] consisting of six histone marks (H3K9ac was not available in both replicates).
- Cell line IMR90 from the ENCODE [66] project consisting of 26 histone marks.

All datasets were available in two replicates and offered the possibility to establish a ground truth for the purpose of validating chromatin state segmentations. A good segmentation method should yield the same or similar chromatin states for two replicate datasets. Furthermore, we can assess the quality of differential regions by looking at the read count fold change between replicates. We expect the read count fold change to be high in differential regions, and low in concordant regions.

Methods were run at binsize 1000 and default parameter settings of each method. We ran `chromstaR` in the modes depicted in SI-Figure 4-19 and in full mode, where all samples are analyzed simultaneously. For `chromstaR` in combinatorial mode (SI-Figure 4-19|b) we set a posterior cutoff of 0.99, as for all analysis in Application 1 in the main text. `jMOSAiCS` could not be run on the monocyte data because input data was not available but required.

In all datasets, `jMOSAiCS` and `chromstaR` consider all the combinatorial state space. For the monocyte dataset, `ChromHMM` and `EpiCSeg` were run considering 7 or 16 states, and `Spectacle` was run considering 7 states (16 is not allowed for four marks). We chose 7 states because this is what chromstaR found for the monocyte data, and 16 states because it is the number of possible combinatorial states. For the Hippocampus dataset, `ChromHMM`, `Spectacle` and `EpiCSeg` were run with 18 states (following [27]). Finally, for the IMR90 dataset, `ChromHMM`, `Spectacle` and `EpiCSeg` were run taking into account 25 states. Runtime and memory requirement were measured on a system with 6 cores @ 2.5 GHz.

The concordance of the segmentation (SI-Figure 4-18|c) was calculated as the fraction of bins that are assigned the same chromatin state in both replicates. We expect this number to be close to 1 for a good segmentation method. However, a high value of this number could also be achieved by assigning the same (not biologically meaningful) chromatin state to the entire genome. For this reason we also report the fraction of the most frequent state along with the concordance. To assess the quality of detected differential regions (regions that do not have the same chromatin state in both replicates), we constructed a normalized read count fold change (SI-Figure 4-18|d). This fold change is 1 if differential regions show the same average change in read count as non-differential regions, and > 1 when differential regions show a higher change in read count as non-differential regions.

The fold change is calculated as follows. Let T be the total number of bins. Indices t, r and m denote the t-th bin (t in [1,T]), the r-th replicate (r in [1,2]) and the m-th histone mark.

- We add a pseudocount of 1 and normalize the binned read count $x_{t,r,m}$ to the total number of reads.

$$x'_{t,r,m} = \frac{x_{t,m,r}+1}{\sum_{t=1}^{T}\left(x_{t,r,m}+1\right)} \qquad \text{(eq. 4.11)}$$

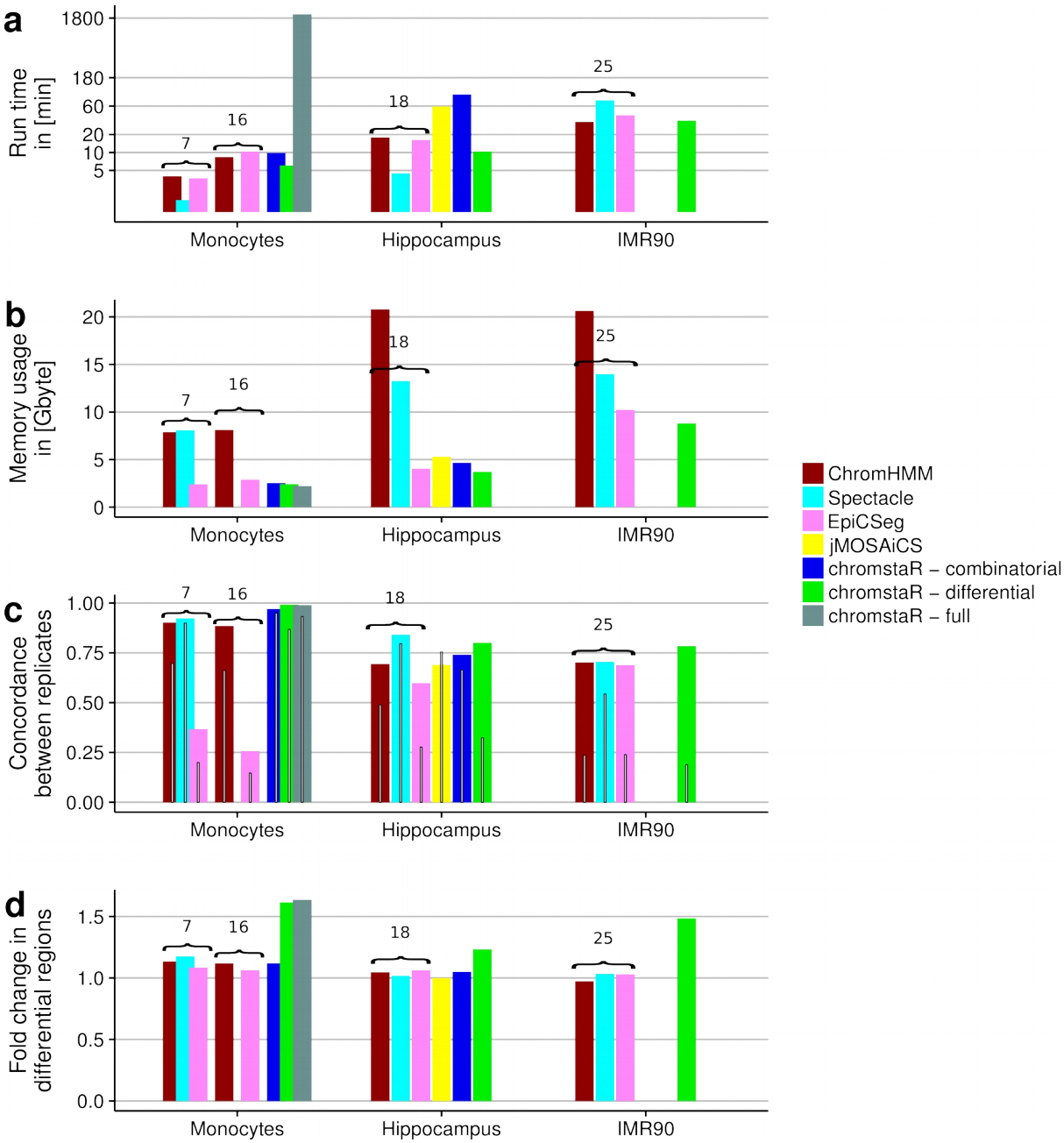
- We calculate a fold change *f* for each bin as

$$f_t = max_m\left(max\left(\frac{x'_{t,1,m}}{x'_{t,2,m}}, \frac{x'_{t,2,m}}{x'_{t,1,m}}\right)\right) \quad \text{(eq. 4.12)}$$

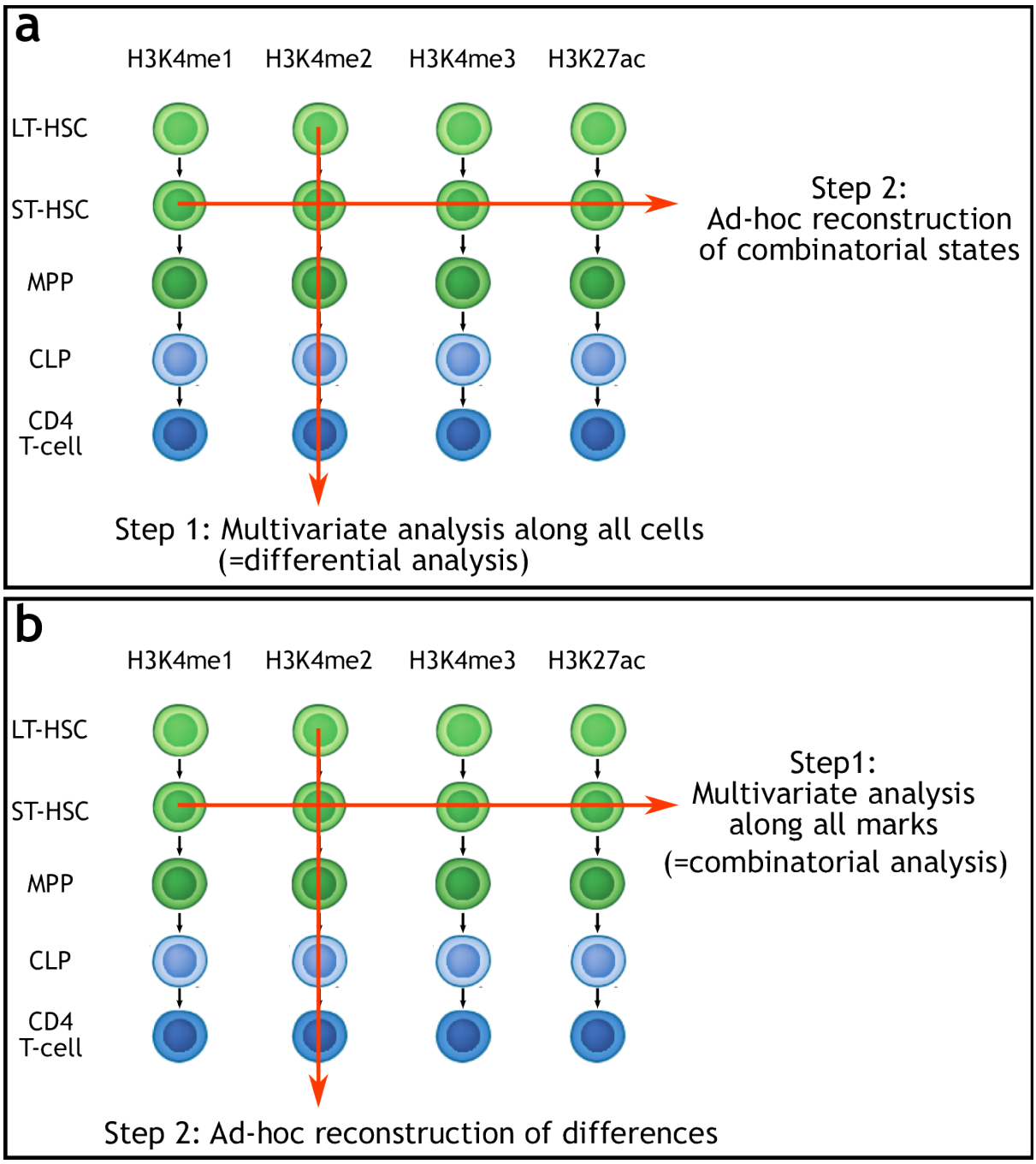
- Finally, we calculate the mean fold change for differential regions and divide it by the mean genomewide fold change. For this last step, we exluced all regions that fall into empty or quiescent states, *i.e.* where the average read count in that state is minimal for all marks. This is necessary for proper normalization, because enriched (non-empty) regions show much higher fold changes than empty regions.

Results of this analysis are depicted in SI-Figure 4-18. We find that chromstaR in differential and full mode shows the most consistent state calls across replicates (SI-Figure 4-18|c), and accurately identifies regions with high read count fold change as differential (SI-Figure 4-18|d). In combinatorial mode, chromstaR performs as well as other multivariate methods. Run time and memory requirement vary a lot for each method depending on the dataset, and no method is clearly the fastest on all datasets.

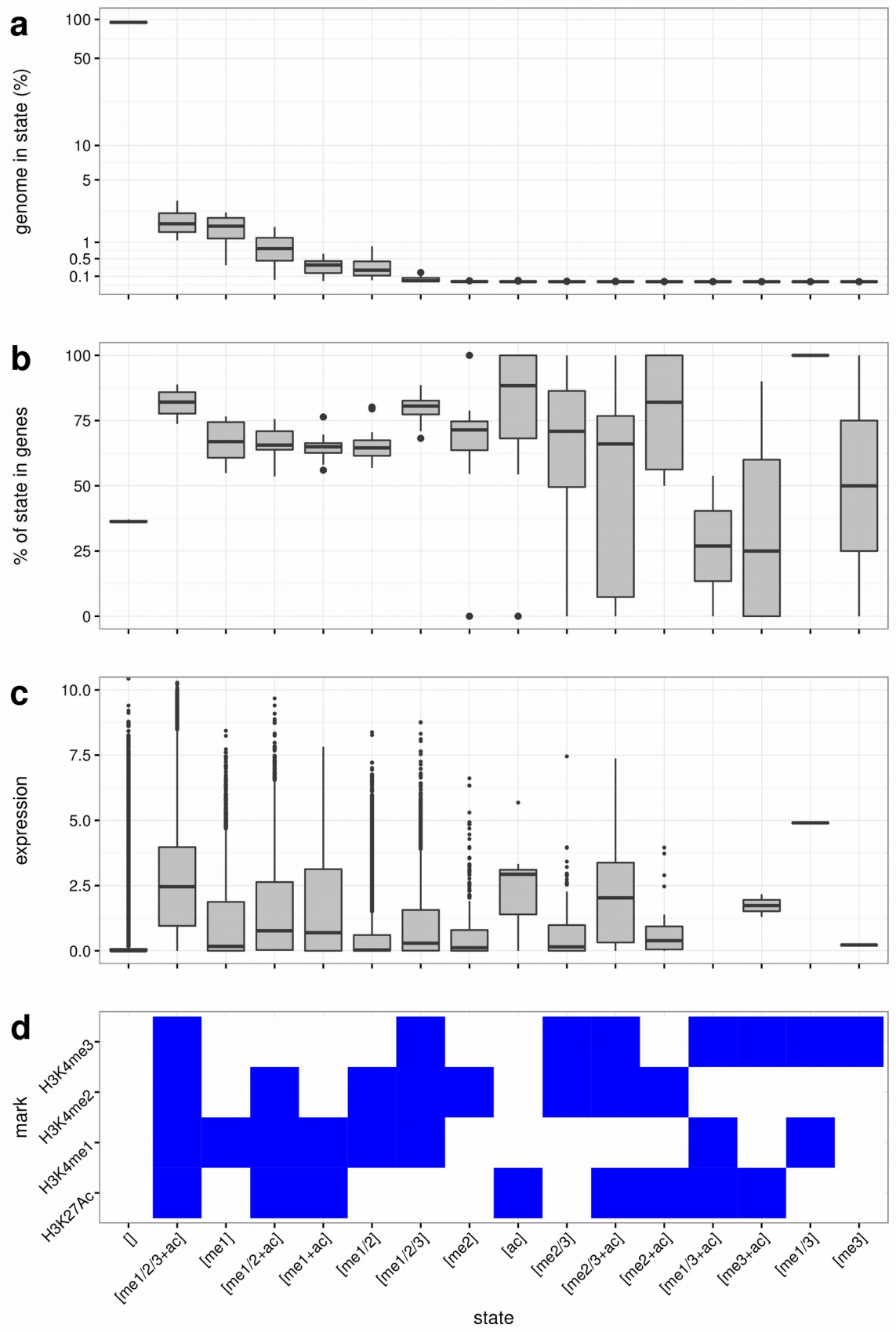
Supplemental Figures



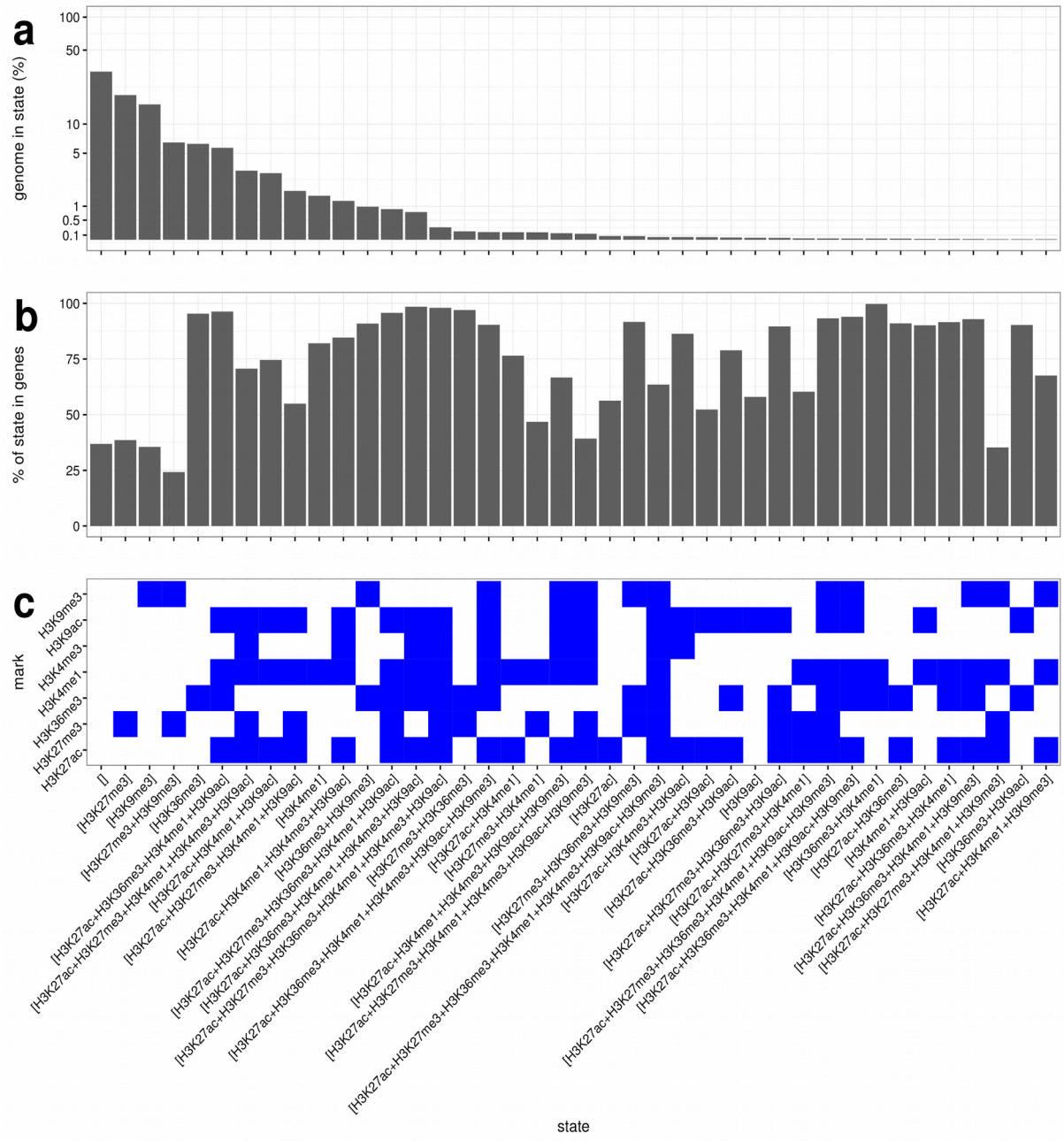
SI-Figure 4-18 | Comparison of multivariate methods. **a** | Run time and **b** | memory requirement for the different methods per dataset with 6 cores @ 2.5~GHz. **c** | The concordance between replicate datasets is calculated as the fraction of bins that have the same chromatin state between replicates. White bars inside the colored bars show the fraction of the genome in the most frequent chromatin state. Panel **d** | shows the normalized fold change in read count for regions that are detected as differential. Missing bars indicate that an algorithm could not be run or finished with error on a particular dataset, due to memory requirements or other reasons. Curly braces specify the number of states which were used to run `ChromHMM`, `Spectacle` and `EpiCSeg`. `jMOSAiCS` and `chromstaR` consider the full combinatorial state space. (Source: Taudt et al. 2016, [51])



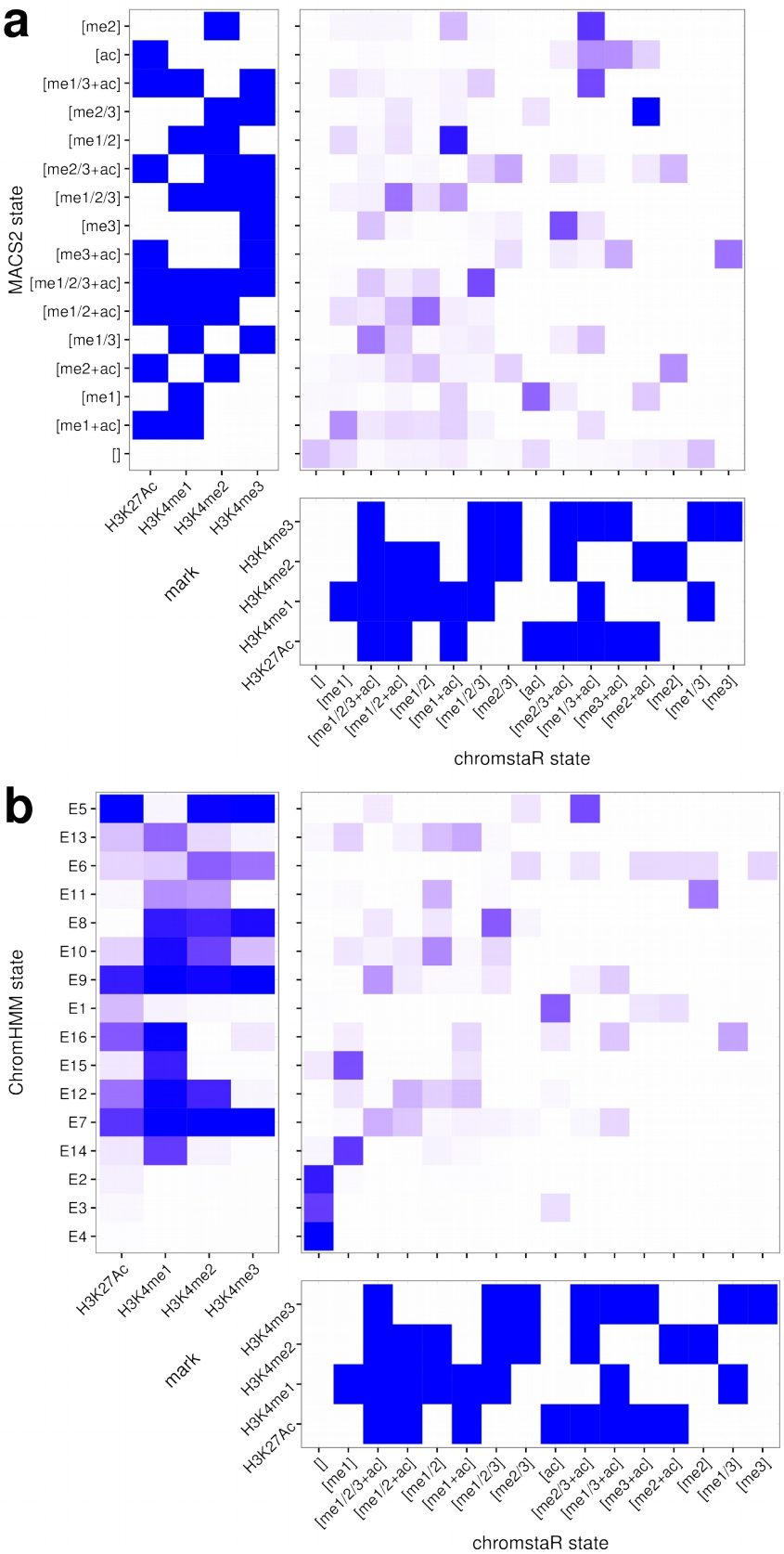
SI-Figure 4-19 | Two-step approach for inferring combinatorial state differences. **a** | In a first step, multivariate peak-calls are obtained along all cells for each mark separately (differential analysis). Those calls are then combined, ad-hoc, into the combinatorial states. **b** | In a first step, combinatorial states are obtained for each cell using the multivariate approach. Differences between those states are then obtained by a simple comparison between cells. (Source: Taudt et al. 2016, [51])



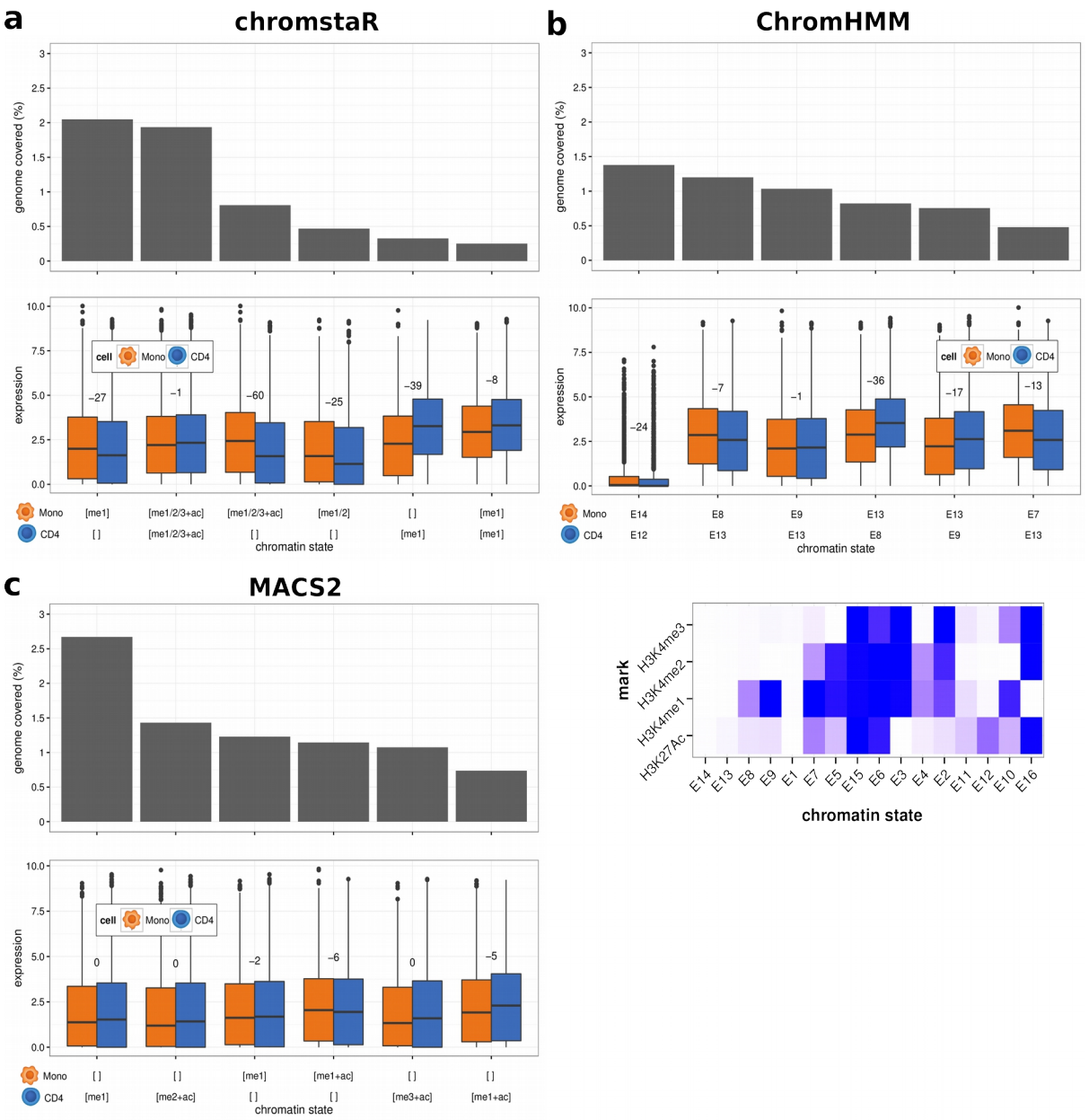
SI-Figure 4-20 | Results for all 16 hematopoietic cell types. **a** | Genomic frequency, *i.e.* the percentage of the genome that is covered by the chromatin state. **b** | Overlap with known genes. **c** | Expression levels of genes whose TSS overlaps the chromatin state. **d** | Heatmap showing the chromatin state definition (blue is present, white is absent). (Source: Taudt et al. 2016, [51])



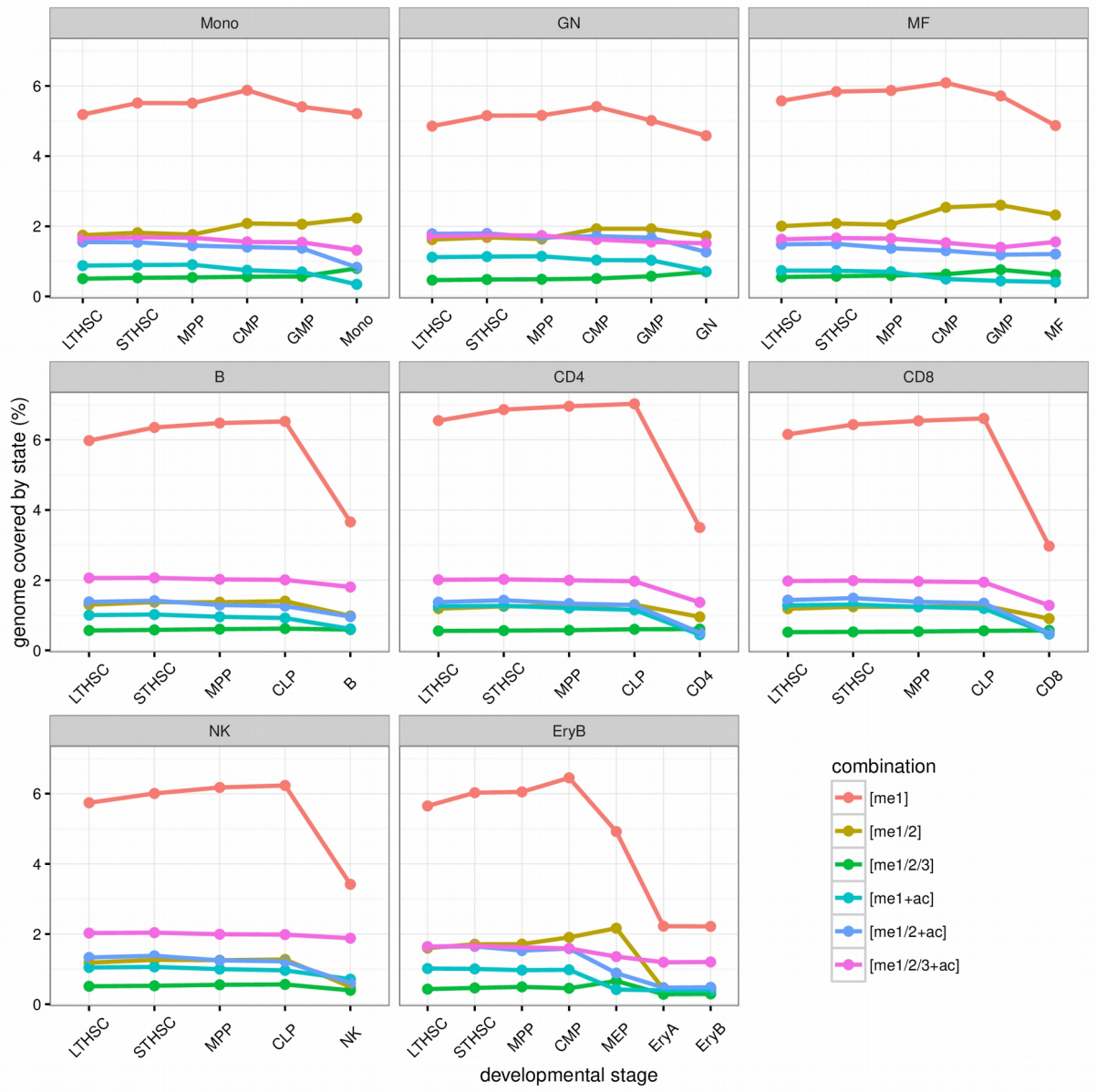
SI-Figure 4-21 | Chromatin states in human Hippocampus tissue. **a** | Genomic frequency, *i.e.* the percentage of the genome that is covered by the chromatin state, for the 40 most frequent states (genomic frequency > 0.01%). **b** | Overlap with known genes. **c** | Heatmap showing the chromatin state definition. (Source: Taudt et al. 2016, [51])



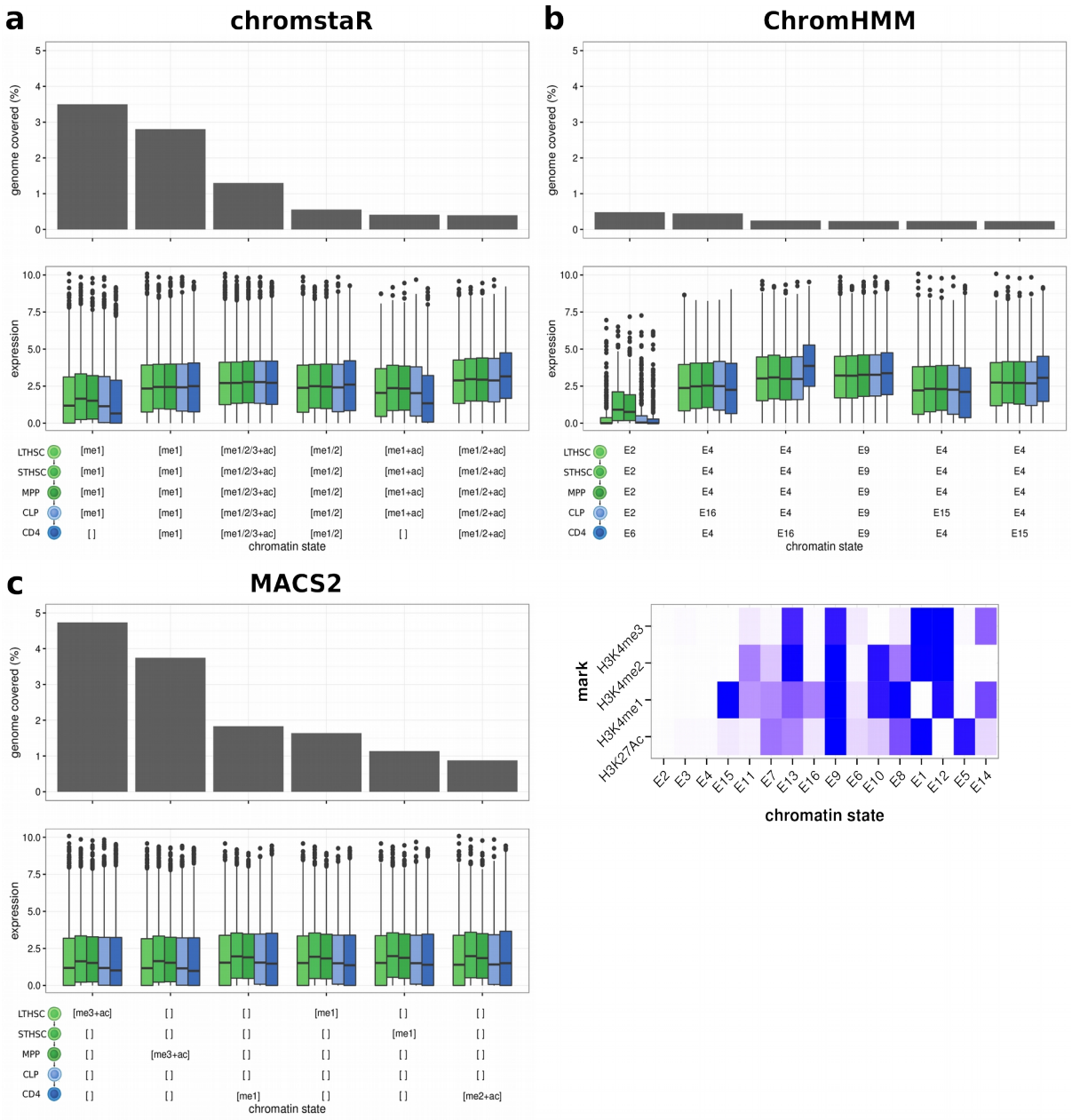
SI-Figure 4-22 | Confusion matrix for the comparison of `chromstaR` with **a** | `MACS2` and **b** | `ChromHMM`. The confusion matrix shows the fold enrichment of states from both methods with each other, with darker tiles (blue) indicating a higher overlap. (Source: Taudt et al. 2016, [51])



SI-Figure 4-23 | Differential analysis of monocytes and CD4 T-cells. Genomic frequency and expression levels for genes that overlap the 6 most frequent differential chromatin states for **a** | `chromstaR`, **b** | `ChromHMM` and **c** | `MACS2`. Numbers give the base-10 logarithm of the multiple testing corrected p-value for the expression difference using a Wilcoxon rank sum test. The more negative the number, the more significant the difference. (Source: Taudt et al. 2016, [51])



SI-Figure 4-24 | Genomic frequency of combinatorial states during differentiation for all branches of the hematopoietic tree (Figure 4-4|a). (Source: Taudt et al. 2016, [51])



SI-Figure 4-25 | Chromatin state transitions for the CD4 branch. Genomic frequency and expression levels for genes that overlap the 6 most frequent chromatin state transitions for **a** | `chromstaR`, **b** | `ChromHMM` and **c** | `MACS2`. (Source: Taudt et al. 2016, [51])

Supplemental Tables

SI-Table 4-3 | The first 10 gene ontology terms for selected differential regions between monocytes and CD4 T-cells after analysis with `chromstaR`. Numbers indicate the binomial false discovery rate (BinomFdrQ) as reported by GREAT. Empty fields indicate that no significant terms were found. (Source: Taudt et al. 2016, [51])

Mono	me1/2/3+ac		me1/2/3+ac	
CD4	me1/2/3+ac		empty	
1	mature B cell differentiation	3.03E-16	response to other organism	5.73E-103
2	apoptotic mitochondrial changes	3.36E-16	response to biotic stimulus	8.46E-103
3	regulation of nuclear-transcribed mRNA catabolic process, ...	3.05E-15	immune response	2.04E-98
4	antigen processing and presentation of exogenous peptide antigen	6.76E-15	inflammatory response	1.00E-80
5	positive regulation of mRNA catabolic process	1.06E-14	response to bacterium	1.66E-74
6	regulation of RNA stability	1.92E-14	positive regulation of immune system process	6.83E-68
7	nuclear-transcribed mRNA catabolic process	2.16E-14	leukocyte activation	5.59E-65
8	regulation of mRNA stability	6.32E-13	regulation of immune response	3.73E-58
9	positive regulation of nuclear-transcribed mRNA catabolic process, ...	1.02E-12	regulation of cytokine production	5.04E-58
10	GPI anchor metabolic process	1.19E-12	response to molecule of bacterial origin	1.30E-57

Mono	empty		me1	
CD4	me1/2/3+ac		me1	
1	leukocyte activation	5.04E-58	regulation of immune system process	1.19E-57
2	T cell activation	3.75E-55	leukocyte activation	5.19E-56
3	T cell differentiation	6.30E-54	cell activation	7.78E-54
4	regulation of immune system process	6.49E-54	immune system development	1.64E-48
5	cell activation	6.41E-53	lymphocyte activation	2.05E-48
6	lymphocyte activation	1.09E-52	hematopoietic or lymphoid organ development	2.42E-46
7	lymphocyte differentiation	8.75E-51	hemopoiesis	4.52E-40
8	alpha-beta T cell activation	1.88E-49	positive regulation of immune system process	1.42E-38
9	immune system process	2.00E-46	lymphocyte differentiation	1.45E-37
10	leukocyte differentiation	3.29E-46	regulation of lymphocyte activation	8.45E-35

Mono	me1		empty	
CD4	empty		me1	
1	response to other organism	5.14E-186	regulation of immune system process	9.31E-114
2	response to bacterium	6.91E-138	leukocyte activation	1.14E-109
3	inflammatory response	4.14E-129	lymphocyte activation	7.68E-99
4	response to molecule of bacterial origin	1.99E-116	cell activation	3.07E-98
5	immune effector process	6.18E-106	T cell activation	7.44E-97
6	response to lipopolysaccharide	5.30E-98	lymphocyte differentiation	3.73E-88
7	negative regulation of transferase activity	9.94E-96	immune system process	6.56E-87
8	myeloid cell differentiation	1.80E-93	leukocyte differentiation	1.11E-81
9	negative regulation of kinase activity	5.11E-90	T cell differentiation	5.16E-81
10	homeostasis of number of cells	4.63E-87	alpha-beta T cell activation	1.06E-77

SI-Table 4-4 | The first 10 gene ontology terms for selected differential regions between monocytes and CD4 T-cells after analysis with `MACS2`. Numbers indicate the binomial false discovery rate (BinomFdrQ) as reported by GREAT. Empty fields indicate that no significant terms were found. (Source: Taudt et al. 2016, [51])

Mono	empty	empty	me1	me1+ac		empty	empty	
CD4	me1	me2+ac	empty	empty		me3+ac	me1+ac	
1				myeloid cell differentiation	2.53E-74		T cell activation	7.45E-56
2				homeostasis of number of cells	2.08E-66		T cell differentiation	1.34E-42
3				cytokine-mediated signaling pathway	3.12E-57		alpha-beta T cell activation	2.07E-37
4				myeloid cell homeostasis	1.11E-49		alpha-beta T cell differentiation	4.45E-33
5				erythrocyte homeostasis	1.26E-48		T cell activation involved in immune response	9.90E-30
6				erythrocyte differentiation	1.72E-47		leukocyte activation involved in immune response	1.32E-23
7				B cell differentiation	4.06E-45		lymphocyte activation involved in immune response	1.64E-21
8				Ras protein signal transduction	9.96E-40		T cell selection	4.14E-20
9				myeloid leukocyte differentiation	5.32E-37		T cell proliferation	6.78E-16

SI-Table 4-5 | The first 10 gene ontology terms for selected differential regions between monocytes and CD4 T-cells after analysis with ChromHMM. Numbers indicate the binomial false discovery rate (BinomFdrQ) as reported by GREAT. Empty fields indicate that no significant terms were found. (Source: Taudt et al. 2016, [51])

Mono	E14		E8	
CD4	E12		E13	
1	homophilic cell adhesion	1.55E-245	immune system process	6.79E-181
2	neuron recognition	3.62E-73	response to biotic stimulus	2.71E-105
3	axon choice point recognition	9.24E-59	response to other organism	7.32E-105
4	axon midline choice point recognition	6.50E-55	leukocyte activation	6.98E-100
5	negative chemotaxis	1.29E-42	immune response	1.15E-95
6	startle response	8.40E-40	cell activation	1.90E-92
7	olfactory bulb interneuron differentiation	4.42E-38	immune system development	3.17E-87
8	innervation	5.91E-25	positive regulation of immune system process	1.59E-77
9	gamma-aminobutyric acid signaling pathway	1.68E-16	response to bacterium	2.96E-74
10	corticospinal tract morphogenesis	6.83E-16	immune effector process	1.67E-69

Mono	E9		E13	
CD4	E13		E8	
1	intrinsic apoptotic signaling pathway	1.06E-39	leukocyte activation	1.18E-60
2	apoptotic mitochondrial changes	2.48E-35	cell activation	8.88E-59
3	myeloid cell homeostasis	1.23E-31	hematopoietic or lymphoid organ development	6.70E-55
4	nuclear export	1.97E-31	chromatin modification	3.07E-54
5	B cell differentiation	6.51E-30	protein modification by small protein conjugation or removal	4.73E-54
6	regulation of intrinsic apoptotic signaling pathway	2.14E-26	immune system development	8.33E-54
7	response to starvation	9.75E-26	lymphocyte activation	1.15E-52
8	regulation of mitochondrial membrane permeability	1.73E-25	protein modification by small protein conjugation	6.90E-52
9	fatty acid biosynthetic process	7.16E-25	hemopoiesis	8.65E-51
10	erythrocyte homeostasis	1.13E-24	protein ubiquitination	7.79E-47

Mono	E13		E7	
CD4	E9		E13	
1	T cell activation	6.61E-53	immune system process	1.10E-168
2	intra-Golgi vesicle-mediated transport	3.76E-40	response to biotic stimulus	1.65E-104
3	peptidyl-lysine modification	1.75E-39	response to other organism	1.19E-103
4	macromolecule methylation	1.54E-37	regulation of immune system process	6.01E-93
5	protein methylation	1.65E-37	leukocyte activation	2.22E-92
6	protein acylation	3.38E-36	multi-organism process	1.37E-91
7	protein acetylation	1.44E-34	cell activation	1.55E-90
8	regulation of B cell activation	1.24E-32	immune response	5.45E-90
9	internal protein amino acid acetylation	1.42E-32	response to bacterium	7.55E-74
10	ncRNA processing	8.95E-32	myeloid leukocyte activation	1.53E-73

SI-Table 4-6 | The first 10 gene ontology terms for selected regions in the CD4 T-cells lineage after analysis with `chromstaR`. Numbers indicate the binomial false discovery rate (BinomFdrQ) as reported by `GREAT`. Empty fields indicate that no significant terms were found. (Source: Taudt et al. 2016, [51])

LTHSC	me1		me1	
STHSC	me1		me1	
MPP	me1		me1	
CLP	me1		me1	
CD4	empty		me1	
1	negative regulation of MAP kinase activity	1.50E-47	apoptotic signaling pathway	4.97E-119
2	myeloid leukocyte activation	6.37E-47	lymphocyte differentiation	4.20E-103
3	phagocytosis	6.23E-42	T cell activation	8.75E-99
4	regulation of lipid kinase activity	7.77E-38	intrinsic apoptotic signaling pathway	3.78E-95
5	cellular response to vascular endothelial growth factor stimulus	9.79E-38	immune effector process	3.37E-92
6	regulation of phosphatidylinositol 3-kinase activity	7.04E-37	regulation of T cell activation	1.67E-87
7	regulation of myeloid leukocyte mediated immunity	9.76E-37	negative regulation of protein kinase activity	3.79E-70
8	negative regulation of multi-organism process	2.01E-34	negative regulation of kinase activity	3.99E-69
9	regulation of lipopolysaccharide-mediated signaling pathway	7.17E-34	negative regulation of transferase activity	8.47E-69
10	regulation of leukocyte degranulation	3.57E-33	regulation of B cell activation	5.50E-67

LTHSC	empty		empty	
STHSC	me1		empty	
MPP	me1		empty	
CLP	me1		empty	
CD4	empty		me1	
1	cell chemotaxis	1.73E-16	leukocyte activation	3.08E-34
2	leukocyte migration	3.96E-14	alpha-beta T cell activation	5.58E-33
3	mesodermal cell differentiation	4.88E-14	T cell activation	9.09E-31
4	positive regulation of peptidyl-tyrosine phosphorylation	5.22E-14	regulation of immune system process	1.06E-30
5	leukocyte chemotaxis	1.28E-13	leukocyte differentiation	1.72E-30
6	platelet-derived growth factor receptor signaling pathway	2.10E-12	lymphocyte differentiation	5.32E-30
7	regulation of vascular endothelial growth factor receptor signaling pathway	8.72E-12	lymphocyte activation	6.06E-30
8	odontogenesis of dentin-containing tooth	4.57E-11	cell activation	1.39E-29
9	positive regulation of cell migration involved in sprouting angiogenesis	1.76E-10	T cell differentiation	2.57E-26
10	glomerulus vasculature development	2.90E-10	T cell selection	1.37E-25

LTHSC	me1	
STHSC	me1	
MPP	empty	
CLP	empty	
CD4	empty	
1	regulation of mRNA stability	1.17E-05
2	regulation of small GTPase mediated signal transduction	2.17E-04
3	retinoic acid receptor signaling pathway	2.30E-04
4	myeloid cell differentiation	4.14E-04
5	microtubule-based transport	9.29E-04
6	negative regulation of lymphocyte proliferation	1.02E-03
7	cytoskeleton-dependent intracellular transport	1.07E-03
8	otolith development	1.22E-03
9	negative regulation of leukocyte proliferation	1.22E-03
10	otolith morphogenesis	1.47E-03

SI-Table 4-7 | The first 10 gene ontology terms for selected regions in the CD4 T-cells lineage after analysis with `chromstaR`. Numbers indicate the binomial false discovery rate (BinomFdrQ) as reported by `GREAT`. Empty fields indicate that no significant terms were found. (Source: Taudt et al. 2016, [51])

LTHSC	me1/2/3+ac		me1/2/3+ac	
STHSC	me1/2/3+ac		me1/2/3+ac	
MPP	me1/2/3+ac		me1/2/3+ac	
CLP	me1/2/3+ac		me1/2/3+ac	
CD4	me1/2/3+ac		empty	
1	protein folding	5.02E-17	immune system process	3.36E-38
2	GPI anchor metabolic process	5.57E-14	immune response	1.81E-20
3	GPI anchor biosynthetic process	9.14E-14	homeostasis of number of cells	1.30E-16
4	apoptotic mitochondrial changes	3.73E-13	immune system development	2.61E-15
5	nuclear export	1.41E-11	regulation of immune response	5.73E-15
6	intrinsic apoptotic signaling pathway in response to DNA damage	5.05E-11	hematopoietic or lymphoid organ development	1.48E-14
7	protein lipidation	2.45E-10	B cell activation	8.12E-14
8	positive regulation of mRNA catabolic process	5.04E-10	leukocyte activation	1.46E-13
9	vacuole organization	1.08E-08	hemopoiesis	2.56E-13
10	regulation of nuclear-transcribed mRNA catabolic process, …	1.23E-08	myeloid leukocyte differentiation	3.35E-13

LTHSC	empty		me1/2/3+ac	me1/2/3+ac
STHSC	empty		me1/2/3+ac	me1/2/3+ac
MPP	empty		empty	empty
CLP	empty		empty	empty
CD4	me1/2/3+ac		empty	me1/2/3+ac
1	T cell differentiation	9.52E-06		
2	lymphocyte differentiation	1.25E-05		
3	lymphocyte activation	1.39E-05		
4	immune system process	1.44E-05		
5	T cell activation	1.54E-05		
6	T cell selection	2.04E-05		
7	T cell receptor V(D)J recombination	2.06E-05		
8	leukocyte activation	2.39E-05		
9	positive T cell selection	2.40E-05		
10	leukocyte differentiation	4.12E-05		

Chapter 5

METHimpute: Imputation-guided construction of saturated methylomes from WGBS data

Aaron Taudt[1,2], David Roquis[3], Amaryllis Vidalis[3], René Wardenaar[3], Frank Johannes[3], Maria Colomé-Tatché[1,2]

1. *European Research Institute for the Biology of Ageing, University of Groningen, University Medical Center Groningen, A. Deusinglaan 1, Groningen 9713 AV, The Netherlands.*
2. *Institute of Computational Biology, Helmholtz Zentrum München, Ingolstädter Landstr. 1, Neuherberg 85764, Germany.*
3. *Department of Plant Sciences, Hans Eisenmann-Zentrum for Agricultural Sciences, Technical University Munich, Liesel-Beckmann-Str. 2, 85354 Freising, Germany.*

Adapted from BMC Genomics 2018; doi: 10.1186/s12864-018-4641-x

Abstract

Whole genome bisulfite sequencing (WGBS) has become the standard method for interrogating plant methylomes at base resolution. However, deep WGBS measurements remain cost prohibitive for large, complex genomes and for population-level studies. As a result, most published plant methylomes are sequenced far below saturation, with a large proportion of cytosines having either missing data or insufficient coverage. Here we present METHimpute, a Hidden Markov Model (HMM) based imputation algorithm for the analysis of WGBS data. Unlike existing methods, METHimpute enables the reconstruction of complete methylomes by inferring the methylation status and level of all cytosines in the genome regardless of coverage. Application of METHimpute to maize, rice and Arabidopsis shows that the algorithm infers cytosine-resolution methylomes with high accuracy from data as low as 6 X, compared to data with 60 X, thus making it a cost-effective solution for large-scale studies. Although METHimpute has been extensively tested in plants, it should be broadly applicable to other species.

Introduction

Cytosine methylation (5mC) is a widely conserved epigenetic mark [98]–[101] with important roles in the regulation of gene expression and the silencing of transposable elements (TEs) and repeats [102], [103]. Experimentally-induced changes in 5mC patterns have been shown to affect plant phenotypes [104]–[106], rates of meiotic recombination [107]–[110], genome stability [111]–[115] and alter plant-environment interactions [116]–[119]. Similar to genetic mutations, changes in 5mC patterns can also occur spontaneously as a result of errors in DNA methylation maintenance [76], [120]–[122]. There is substantial evidence in plants that experimentally-induced as well as spontaneously occurring 5mC changes can be stably inherited across multiple generations, independently of genetic changes [123]. Cytosine methylation has therefore emerged as a potentially important factor in plant evolution [124]–[126] and as a possible molecular target for the improvement of commercial crops [127], [128].

Plant methylomes are now routinely studied using whole genome bisulfite sequencing (WGBS), a next generation sequencing (NGS) method that can interrogate the methylation status of individual cytosines at the genome-wide scale. The application of this technology has been instrumental in dissecting the molecular pathways that establish and maintain 5mC patterns in plant genomes. Unlike in animals, plants methylate cytosines in context CG, but also extensively in contexts CHG and CHH, where H = A, T, C [102]. Methylation at CG dinucleotides (mCG) is maintained by methyltransferase 1 (MET1), which is recruited to hemi-methylated CG sites in order to methylate the complimentary strand in a template-dependent manner during DNA replication [129]. By contrast, mCHG is maintained dynamically by the plant specific chromomethylase 3 (CMT3) [130], and requires continuous interactions with H3K9me2 (dimethylation of lysine 9 on histone 3) [131]. Asymmetrical methylation of CHH sites (mCHH) is established and maintained by another member of the CMT family, CMT2 [99], [132]. Similar to CMT3, CMT2 dynamically methylates CHH in H3K9me2-associated regions. In addition to these context-specific maintenance mechanisms, all three sequence contexts can also be methylated de novo via RNA-directed DNA methylation (RdDM) [102], which involves short-interfering 24 nucleotide small RNAs (siRNA) that guide the de novo methyltransferase domains rearranged methyltransferase 2 (DRM2) to homologous target sites throughout the genome [133], [134].

Although these methylation pathways appear to be broadly conserved across plant species, recent data indicates that there is extensive variation in 5mC patterns both between but also within species [100], [135]. Efforts to explore the origin of this variation and its implications for plant evolution, ecology and agriculture will require large inter- and intraspecific methylome datasets. Such datasets are currently emerging. To date, the methylomes of over 50 plant species have been analyzed using WGBS [100], [101], including representative species of major taxonomic groups such as angiosperms, gymnosperms, ferns, and non-vascular plants. In addition, the methylomes of over 1000 natural *A. thaliana* accessions are now available [136], as well as those of several experimentally derived populations [137]. However, deep inter- and intraspecific WGBS measurements remain cost-prohibitive,

particularly for species with large genomes. Most published plant methylomes have therefore been sequenced far below saturation (*i.e.* large number of cytosines in the genome are not covered). Indeed, even simple genomes, like that of the model plant *A. thaliana* (Col-0 accession), are typically only sequenced to about 10-30 X. At this depth, about 5-10% of cytosines have missing data (*i.e.* zero read coverage) and about 15-20% have nearly uninformative read coverage (< 3 reads), and this problem is exacerbated in more complex genomes, like those of rice and maize (see Figure 5-4).

Low to moderate sequencing depths in individual samples have cumulative consequences for analyzing population-level data. For instance, in the recently released 1000 *A. thaliana* methylome data [136] (measured at 5 X coverage per strand on average), 92% of cytosines have missing data in at least one sample when 100 accessions are compared (Figure 5-1). These incomplete measurements will reduce statistical power in genome-wide methylation QTL (meQTL) mapping studies, in estimates of epimutation rates, or in ecological studies that aim to correlate site-specific methylation levels with environmental/climatic variables. Moreover, incomplete measurements also complicate and potentially bias methylome scans for signature of epigenetic selection using methylation site frequency spectrum (mSFS) analytic approaches [124]. One way to circumvent the missing data problem is to calculate methylation levels over larger regions, ranging anywhere from several hundred to several thousand basepairs and to use these methylation levels for downstream population-level analyses. In the above-mentioned *A. thaliana* population data, only 36% of 100 bp regions in the genome are missing in at least one sample of the 100 accessions, compared with 92% of individual cytosines, and this percentage further decreases with larger region sizes. However, while region-based methylation levels are useful measures for descriptive and correlative analyses, these measures obscure detailed insights into the cytosine-level methylation status calls, and thus arguably undermine the key advantages of WGBS over other lower resolution technologies such as MeDIP-seq. Cytosine-level status calls are needed to be able to apply existing population (epi)genetic models to population methylome data, and to be able to test explicit hypotheses about the evolutionary forces that shape methylome variation patterns within and among species [136].

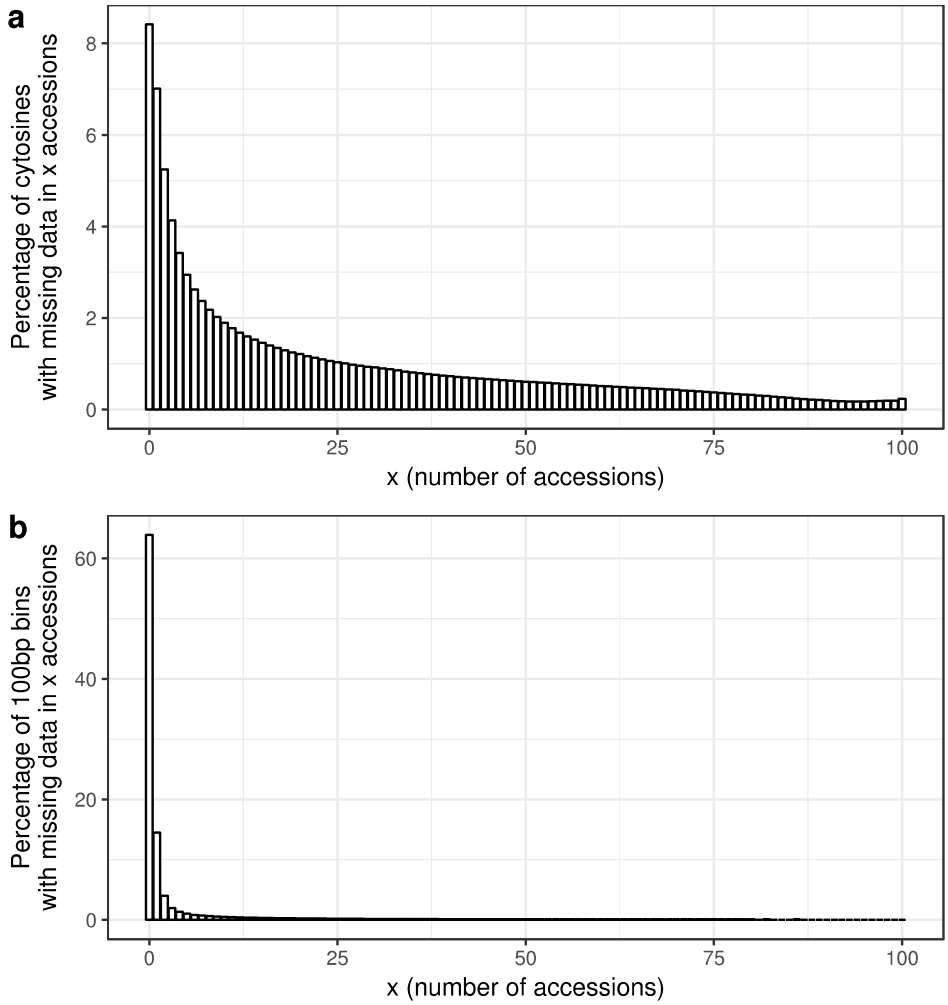


Figure 5-1 | Missing cytosines in population epigenetic studies. For certain applications in population epigenetic studies (*e.g.* meQTL, mSFS, epimutation rates), only positions that are covered in all samples can be used. This leads to substantial dropout of usable positions if the number of samples is high. The y-axis shows the percentage of all **a** | cytosines and **b** | 100 bp bins that are not covered (zero reads) in x samples. For example, in (**a**) ~ 8% of cytosines have missing data in 0 samples, meaning that only 8% of cytosines are covered in all samples, while 92% are missing in at least one sample. The data for this graph is from the 1001 methylomes project [136]. Mean coverage of this study was 5X (per strand and cytosine). (Source: Taudt et al. 2018, [138])

In order to maximize the information contained in WGBS data and to facilitate cost-effective sequencing decisions for future studies, we developed `METHimpute`, a Hidden Markov Model (HMM) based imputation algorithm for the reconstruction of base-resolution methylomes from WGBS data. The unique feature of this algorithm is its ability to impute the methylation status and level of cytosines with missing or uninformative measurements, thus yielding complete methylomes even with low-coverage WGBS datasets. Indeed, using published WGBS data from *Arabidopsis thaliana* (rock cress), *Oryza sativa* (rice) and *Zea mays* (maize), we demonstrate that `METHimpute` accurately reconstructs base-resolution methylomes from data with an average coverage as low as 6 X, suggesting that typical sequencing costs could be cut by more than 50% without a significant loss of information.

Conceptual overview

WGBS is an NGS-based method in which DNA is treated with sodium bisulfite prior to sequencing in order to convert unmethylated cytosines into uracils and finally into thymines during PCR amplification. Hence, a cytosine in a bisulfite treated read that maps to a cytosine in the reference genome provides evidence for methylation, while a thymine that maps to a cytosine does not. Many specialized short read mapping programs make use of this information and output so-called methylation levels [139]–[141]; that is, the proportion of aligned reads that support that a cytosine is methylated out of all the reads covering the site. Methylation levels are inherently noisy due to inefficiencies in the sodium bisulfite conversion step. Moreover, tissue heterogeneity and the highly dynamic maintenance methylation at CHH and CHG, which requires feedback loops with histone modifications and small RNAs [102], [103], lead to intermediate methylation levels which are very susceptible to experimental variation. Finally, in WGBS data a large proportion of cytosines are often either not covered by any sequencing read or are covered only by a few number of reads (Figure 5-4), meaning that methylation levels at these positions cannot be determined.

To overcome these limitations we developed `METHimpute`, a Hidden Markov Model (HMM) for the reconstruction of base-resolution methylomes from WGBS data. `METHimpute` takes methylated and unmethylated read counts at every cytosine as input, and outputs discrete methylation status calls (unmethylated or methylated), together with recalibrated methylation levels between 0 and 1 for every cytosine in the genome, regardless of coverage (Figure 5-2).

The `METHimpute` algorithm fits a two-state HMM to the observed methylation counts. The two hidden states correspond to the unmethylated (U) and methylated (M) components, with component-specific binomial emission densities. The estimates of the binomial parameters (p_U and p_M) and the HMM transition matrix (*i.e.* the collection of probabilities to transition from one hidden state to another) are estimated freely during model training for different sequence contexts, thus requiring no empirical knowledge of the conversion rate. In the present analysis we have used contexts CG, CCG, CWG, CAA, CTA and CCA|CHY (where H={A,C,T}, W={A,T} and Y={C,T}), following evidence of their different methylation characteristics [142]. If necessary the model could be extended to account for different emission and transition parameters for every context and annotation category (genes, TEs, CpG density, etc.), instead of context alone, although this would only be possible for well-annotated genomes.

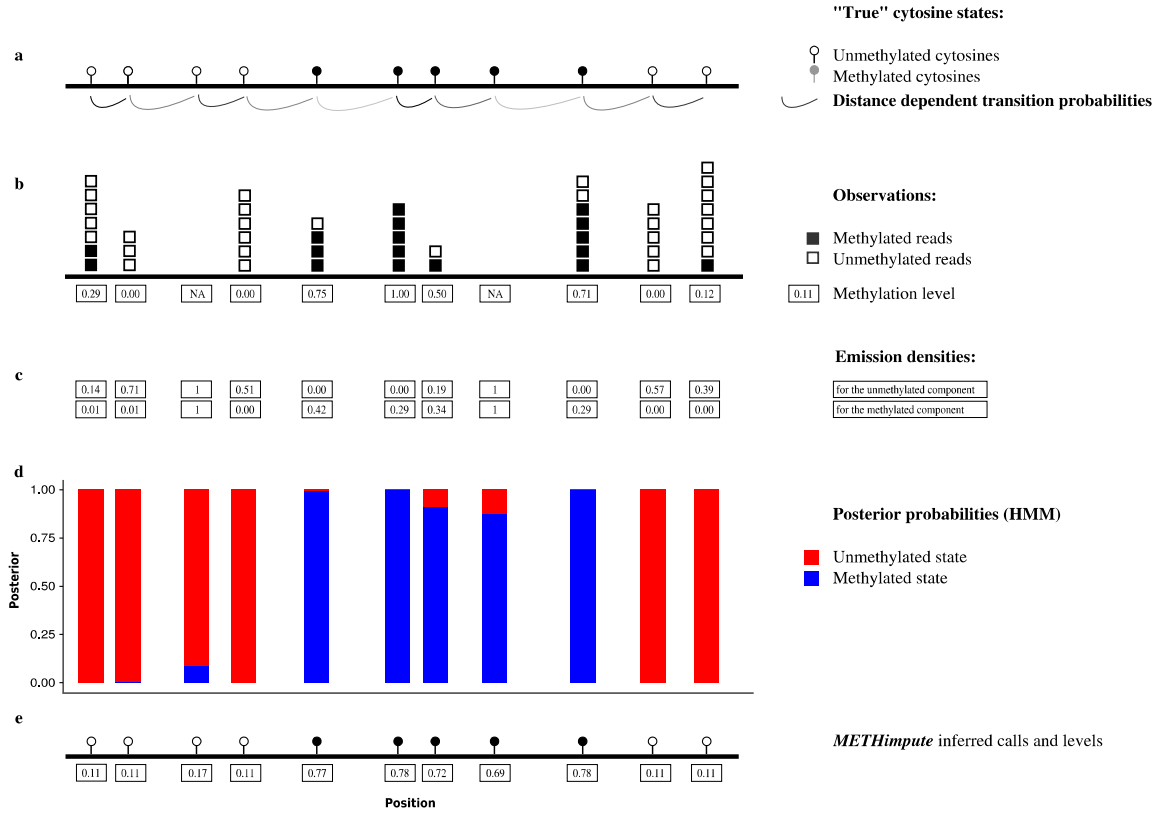


Figure 5-2 | Conceptual overview of METHimpute. **a** | Cytosines on the sequenced genome are assumed to be either unmethylated or methylated. **b** | Bisulphite-sequencing and alignment yields methylation levels for each cytosine, *i.e.* the number of reads showing methylation divided by the total number of reads. **c** | Emission densities for each state are obtained with a binomial test with state-specific parameters. Note that "imputed" cytosines, *i.e.* cytosines without any reads, are treated identically as all other cytosines. However, since the emission densities for all states are 1 for imputed cytosines, the methylation status call is purely driven by the neighborhood of cytosines. **d** | Model fitting yields posterior probabilities for methylation status calls. **e** | Inferred methylation status calls and methylation levels. (Source: Taudt et al. 2018, [138])

Based on the model fits, the probability that a given cytosine belongs to one of the hidden states is given by the posterior probabilities γ_U and γ_M (Figure 5-2|d, Methods section). A cytosine's maximum posterior probability represents its most likely methylation status (Figure 5-2|d,e), and the magnitude of this probability can be used as a measure of confidence in the underlying status call. In addition to methylation status calls, `METHimpute` outputs recalibrated methylation levels per cytosine, calculated as $m' = p_U \cdot \gamma_U + p_M \cdot \gamma_M$ (Figure 5-2|e). A key feature of `METHimpute` is its ability to infer the methylation level and status for cytosines with missing data (*i.e.* zero read coverage) or for those with poor read coverage (*i.e.* less than 3 reads). It achieves this inference iteratively during HMM training by borrowing information from neighboring sites. The algorithm therefore outputs complete, base-resolution methylomes, that can otherwise only be obtained through very high-depth sequencing experiments.

Model specification

Outline

We define an $N = 2$ state Hidden Markov Model (HMM), where the states i represent unmethylated (U) and methylated (M) cytosines. The emission densities for each state are binomial distributions, which can be interpreted as a binomial test on the number of methylated counts m over total counts r. The probability parameter p_i of the binomial test can be interpreted as the probability of finding m methylated counts out of r total counts, given the state i. Note that in this definition 1 - p_U is the conversion rate, *i.e.* the probability of a read showing non-methylation when the cytosine is indeed non-methylated. Cytosines are not equally spaced in the genome, and we therefore chose a distance dependent transition matrix $\boldsymbol{A}$, where the distance dependent change in transition probabilities is modeled by an exponential function. Furthermore, to account for different sequence contexts, we implemented context-specificity for both the binomial test and the transition probabilities.

Mathematical description

The probability P of observing methylated m_t and total r_t read count at a particular cytosine t in context c_t can be written as

$$P_t(m_t, r_t, \boldsymbol{p}_{c_t}) = \sum_{i \in \{U, M\}} \gamma_{it} B_{ic_t}(m_t, r_t, p_{ic_t}) \quad , \qquad \text{(eq. 5.1)}$$

where γ_i are the posteriors (mixing weights) and B_i are binomial distributions with context-specific parameter p_{ic}. The binomial distribution is defined as

$$B(m, r, p) = \binom{r}{m} p^m (1-p)^{r-m} \quad . \qquad \text{(eq. 5.2)}$$

All probability parameters of the binomial tests (*i.e.* the probabilities of a success) are estimated freely during model training (next section). For $C = 6$ contexts and $N = 2$ states, $N \times C = 12$ independent parameters p_{ic} need to be fitted.

The distance dependent transition probabilities from cytosine t in state i to cytosine t+1 in state j, separated by distance $d_{t,\,t+1}$ and in transition context $c_{t,\,t+1}$, can be described as

$$A_{ij,c_{t,t+1}}(A^o_{ij,c_{t,t+1}}, d_{t,t+1}, D_{c_{t,t+1}}, N) = A^o_{ij,c_{t,t+1}} \boldsymbol{e}^{-d_{t,t+1}/D_{c_{t,t+1}}} + \frac{1}{N}\left(1 - \boldsymbol{e}^{-d_{t,t+1}/D_{c_{t,t+1}}}\right) \quad . \qquad \text{(eq. 5.3)}$$

$A^o_{ij,c_{t,t+1}}$ are the transition probabilities without distance dependency (or for adjacent cytosines with $d_{t,\,t+1} = 0$). $D_{c_{t,t+1}}$ is a constant that reflects how fast neighboring cytosines lose correlation. The distance dependency is constructed in such a way that all transitions $A_{ij,c_{t,t+1}}$ are equally likely for an infinite distance $d_{t,\,t+1} = \infty$. Note that for $C = 6$ contexts the model has $C \times C = 36$ transition contexts and thus 36 different transition matrices with dimensions $N \times N$.

The constants D_c are determined by a non-linear least-squares (nls) fit to the correlation decay between cytosines in transition context $c_{t,\,t+1}$ (see Figure 5-3 for all used transition contexts). The formula for the fit is $y_c(d) = a0 * e^{-d/D_c}$, where y_c is the correlation between neighboring cytosines at distance d in transition context c. The parameters $a0$ and D_c are fitted by the nls-fit.

An important point is that the correlation is calculated between adjacent cytosines, with no other cytosines in between. This reflects the definition of the transition probabilities in the Hidden Markov Model, where transitions are defined from one cytosine to the next in the sequence.

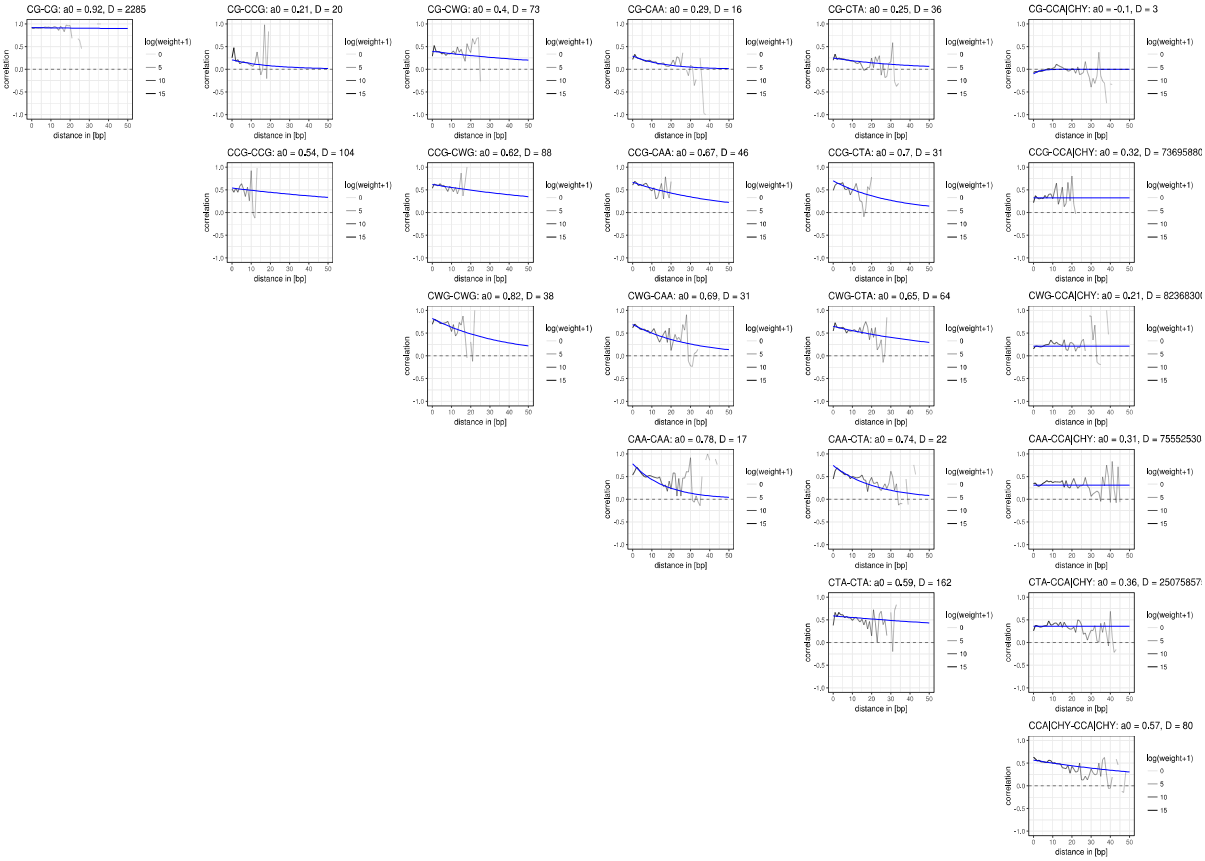


Figure 5-3 | Distance correlation. Correlation between the methylation levels of neighboring cytosines, split by context combinations. The distance is defined as the number of base-pairs in between the two neighboring cytosines (without any other cytosines in between). The blue curve is a weighted exponential fit with formula y = a0 * exp(-x/D). The figure shows correlations from sample “Arabidopsis 8.6X”. (Source: Taudt et al. 2018, [138])

Model fitting

Model parameters are fitted with the Baum-Welch algorithm [1]. The distance-dependent transition probabilities require modified updating formulas compared to a standard Baum-Welch algorithm without distance dependency. The derivation of the modified updating formulas is detailed below, and uses notation introduced in [6]. Please see section "Mathematical notation" in the introduction for details about the notation.

The conditional expectation Q that needs to be maximized can be written as

$$Q = \sum_{i}^{N} \gamma_{i,t=0} \log(\pi_i) + \sum_{i,j,t}^{N,N,T-1} \xi_{ijt} \log(A_{ij,c_{t,t+1}}) + \sum_{i,t}^{N,T} \gamma_{it} \log(B_{it}) \quad \text{. (eq. 5.4)}$$

The updated transition probabilities A'^{o}_{ijc} can be obtained by solving $\frac{\partial Q}{\partial A^{o}_{ijc}} = 0$ using the method of Lagrange multipliers to deal with the constraint $\sum_{j}^{N} A^{o}_{ijc} = 1$.

$$A'^{o}_{ijc} = \left(\sum_{t}^{T-1} \delta_{c,c_{t,t+1}} \xi_{ijt} \frac{A^{o}_{ijc}}{A_{ij,c_{t,t+1}}} \frac{\partial A_{ij,c_{t,t+1}}}{\partial A^{o}_{ij,c_{t,t+1}}} \right) / \left(\sum_{t,j}^{T-1,N} \delta_{c,c_{t,t+1}} \xi_{ijt} \frac{A^{o}_{ijc}}{A_{ij,c_{t,t+1}}} \frac{\partial A_{ij,c_{t,t+1}}}{\partial A^{o}_{ij,c_{t,t+1}}} \right) \text{ .(eq. 5.5)}$$

Here, $\delta_{c,c_{t,t+1}}$ is the Kronecker delta function, which ensures that only terms in the correct transition context c are included into the sum.

Similarly, the updated parameters for the binomial test can be obtained by solving $\frac{\partial Q}{\partial p_{ic}} = 0$. For independent binomial tests, this yields

$$p'_{ic} = \left(\sum_{t}^{T} \delta_{c,c_t} \gamma_{it} m_t \right) / \left(\sum_{t}^{T} \delta_{c,c_t} \gamma_{it} r_t \right) \quad \text{. (eq. 5.6)}$$

The methylation status i_t is determined by maximizing over the posterior probabilities $i_t = \text{argmax}_i(\gamma_{it})$.

Finally, we can use the posterior probabilities $\gamma_{U|M,t}$ and estimated parameters p_{ic} to define a recalibrated methylation level m'_t that is defined on every cytosine t in the genome and can serve as input for other applications:

$$m'_t = p_{U,c_t} \cdot \gamma_{U,t} + p_{M,c_t} \cdot \gamma_{M,t} \quad \text{(eq. 5.7)}$$

Data preparation

We used published data (fastq files containing bisulfite sequencing reads) from three model plant species to test METHimpute: *Arabidopsis thaliana*, rice (*Oryza sativa Japonica cv. Nipponbare*) and maize (*Zea mays* B73). We used three replicates for rice and maize, and two replicates for Arabidopsis. Each sample was mapped to the latest available version of the reference genome for this species. Details and references on these datasets, reference genomes and annotations files, as well as additional alignment metrics can be accessed in SI-Table 2 (online).

Read sequences (SI-Table 2, online) were quality trimmed and adapter sequences were removed with Cutadapt (version 1.9; python version 2.7.9; [143]). Trimming was performed on both ends using the single-end mode and the quality threshold was set to a phred score of 20 (q = 20). We applied the default error rate of 10% for the removal of the adapter sequences. Afterwards, we discarded reads shorter than 40 base pairs. Reads were subsequently mapped to an indexed genome. The maximum allowed proportion of mismatches was set to 0.05 (m = 0.05, 5 mismatches per 100 bp) and the maximum insert size was set to 1000 bp (X = 1000). BS-Seeker2 (v2.0.10; [141]) using Bowtie2 (version 2.2.2; [144]) was chosen for the alignment of the reads. Samtools (version 1.3.1; using htslib 1.2.1; [145]) was used to remove duplicates (samtools rmdup -s) and to sort bam files (samtools sort). Methylomes were subsequently constructed through the bs_seeker2-call_methylation.py module from BS-Seeker2 (v2.0.10; [141]). CGmap files containing methylome information were used as an input for METHimpute.

Results

Imputation-guided reconstruction of complete Arabidopsis, rice and maize methylomes

To demonstrate the performance of METHimpute we analyzed representative WGBS datasets from *A. thaliana* (Col-0) [137], *O. sativa* (japonica nipponbare) [146], and *Z. mays* (B73) [147]. We chose these three species because they cover a wide spectrum of plant genomes in terms of length and complexity: the *A. thaliana*, *O. sativa* and *Z. mays* genomes are 120 Mb, 374 Mb and 2.1 Gb in length, respectively, and have an estimated repeat content of 10%, 28-35% and 85% [148]–[151]. The *A. thaliana* data consisted of two replicates (rep.1: 8.6 X; rep.2: 15.7 X coverage per cytosine per strand), while there were three replicates both for *O. sativa* (rep.1: 7.4 X, rep.2.: 6.9 X, rep.3: 4.6 X) and *Z. mays* (rep.1: 1.6 X, rep.2: 3.3 X, rep.3: 2.4 X). The precise mapping statistics for each dataset are detailed in SI-Table 1 (online). Alignment and pre-processing of the data was carried out using a single pipeline as described in the Methods section. Runtimes and memory requirements for METHimpute are listed in SI-Table 4 (online).

We examined the genome-wide coverage distributions of each replicate dataset. Despite average coverage being relatively high, a substantial proportion of cytosines had either missing data or low coverage. For instance, in the *A. thaliana* (rep.1: 8.6 X), *O. sativa* (rep.3: 4.6 X) and *Z. mays* (rep.2: 3.3 X) datasets, about 9% (3.71M), 24% (39.54M) and 26% (36.77M) of all cytosines had missing data (*i.e.* zero read coverage) and 24% (10.27M), 49% (79.38M) and 60% (85.5M) were nearly uninformative (here defined as coverage < 3 reads) (Figure 5-4|d-f and SI-Figure 5-9 for the other replicates). Interestingly, the genome-wide proportions of missing or uninformative sites were highly context dependent, being highest for CCA|CHY, probably as a result of less unique short read alignments in this context as it is more abundant in repetitive regions of the genome (SI-Figure 5-10 and Figure 5-7).

We applied `METHimpute` to the above-described datasets and evaluated the quality of the resulting methylation calls. For *A. thaliana*, *O. sativa* and *Z. mays*, the algorithm imputed the methylation status of all 3.71M, 39.54M and 36.77M missing data cytosines, respectively, and inferred the methylation status of all 10.27M, 79.38M and 85.5M uninformative cytosines.

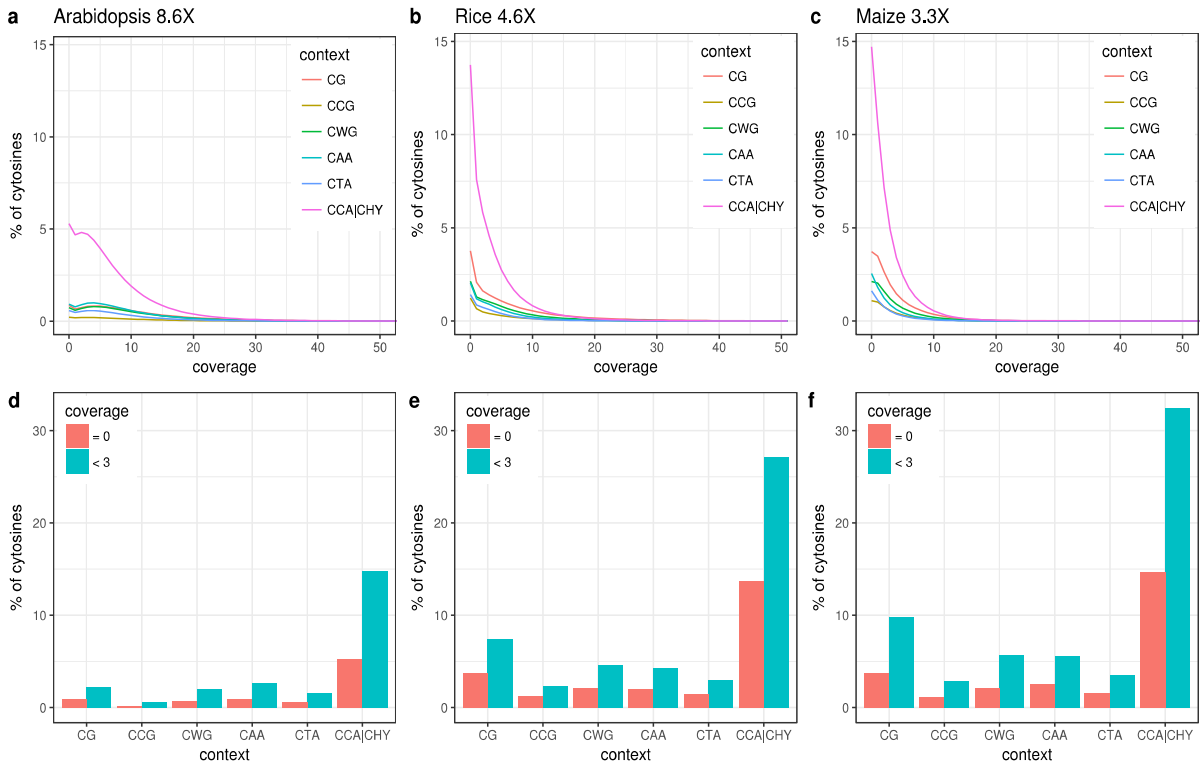


Figure 5-4 | Coverage distributions. **a-c** | Percentage of cytosines with X coverage (strand-specific). **d-f** | Percentage of cytosines with missing data (red) and "uninformative" coverage (green), defined as less than three reads. (Source: Taudt et al. 2018, [138])

Inferred methylation calls capture known biology

To evaluate the quality of the inferred methylation status calls and levels we examined the per-cytosine posterior probability of being either unmethylated (U) or methylated (M). As mentioned above, this probability represents a measure of statistical confidence in the underlying methylation call, with a value of 1 being the most confident. We found that the distribution of maximum posterior probability values for imputed cytosines shows a clear peak around 1 and a tail of lower confidence values (Figure 5-5 and SI-Figure 5-11 for the other replicates), suggesting that the algorithm produces high-confidence methylation calls for a large proportion of cytosines with missing data. Indeed, 58% (1.50M), 54% (3.96M) and 83% (6.43M) of imputed cytosines in *A. thaliana*, *O. sativa* and *Z. mays* were called with high confidence (defined as posterior probability ≥ 0.9), and these numbers increased to 91% (4.16M), 90% (6.64M) and 93% (9.56M) for cytosines covered by only one or two reads.

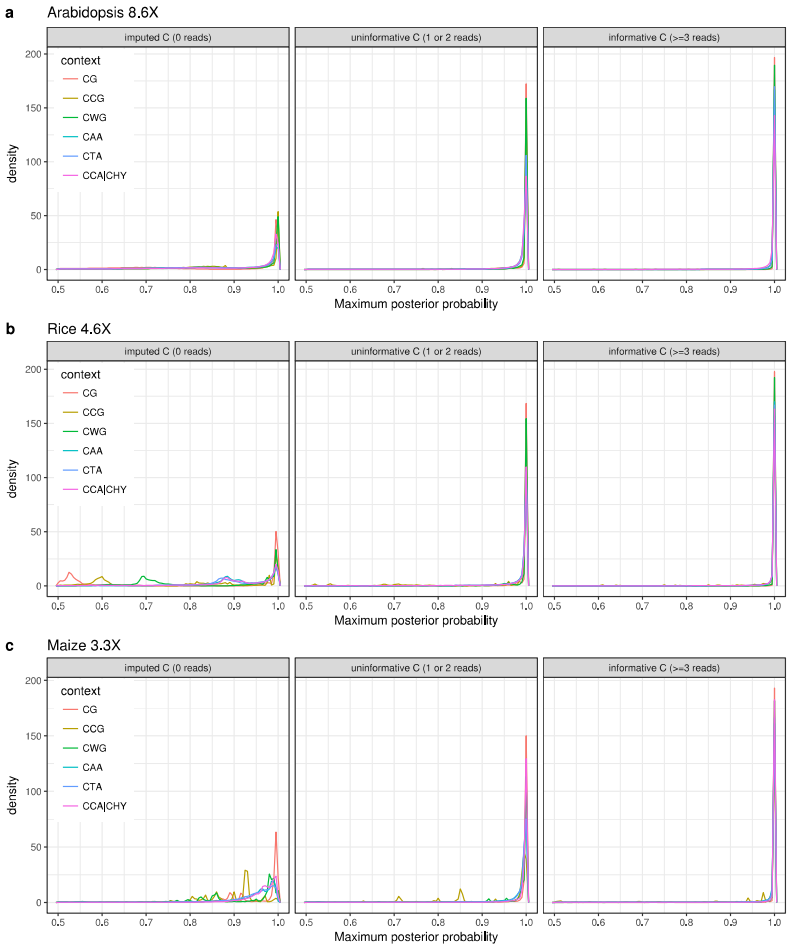


Figure 5-5 | Maximum posterior distributions for imputed cytosines (coverage = 0), uninformative cytosines (coverage = 1 or 2) and informative cytosines (coverage ≥ 3). The figure shows the distributions of the maximum posterior probabilities with density on the y-axis and the maximum posterior probability on x-axis. The maximum posterior probability, *i.e.* the confidence in the methylation status calls, is generally lower for sites with less coverage. (Source: Taudt et al. 2018, [138])

To assess whether the inferred methylation levels are consistent with known biology, we constructed meta-methylation profiles for annotated repeats and genes using cytosines separated in three different categories: informative (coverage ≥ 3), uninformative (coverage = 1 or 2) and imputed cytosines (coverage = 0). Regardless of coverage category, `METHimpute` confirms that *A. thaliana* TE sequences are heavily methylated in all sequence contexts, with a marked decrease in methylation levels at their 5' and 3' ends (Figure 5-6|b and SI-Figure 5-12|b for the other replicate). The CCA|CHY context shows the lowest methylation levels and CG shows the highest, consistent with Gouil and Baulcombe [142], and the ordering is conserved for imputed and uninformative cytosines. Similar profiles were detected for repeat elements in *O. sativa* and *Z. mays*, with high CG, CCG and CWG methylation, and very low levels of CAA, CTA, and particularly CCA|CHY methylation, consistent with known results (Figure 5-6|d,f and SI-Figure 5-12 for the other replicates) [152].

In line with numerous methylome studies in Arabidopsis (*e.g.* [142], [153], [154]), `METHimpute` finds that *A. thaliana* genes are intermediately methylated in CG context, and essentially unmethylated at all CHG (CCG, CWG) and CHH (CAA, CTA, CCA|CHY) sites (Figure 5-6|a and SI-Figure 5-12|a for the other replicate). Genic meta-methylation profiles for *O. sativa* and *Z. mays* were generally similar to those of *A. thaliana* (Figure 5-6|c,e and SI-Figure 5-12 for the other replicates), with the exception that both crop species are known to also methylate genic CHG context, probably owing to the fact that genes in these complex genomes often overlap or contain heavily methylated TE or repeat copies.

Taken together the above analyses illustrate two points: first, `METHimpute` infers annotation-specific methylation profiles that are consistent with published reports; and second, the methylation profiles inferred from imputed or uninformative cytosines recapitulate the patterns seen for highly-informative cytosines, indicating that – regardless of coverage – the inferred methylation calls are robust and biologically meaningful.

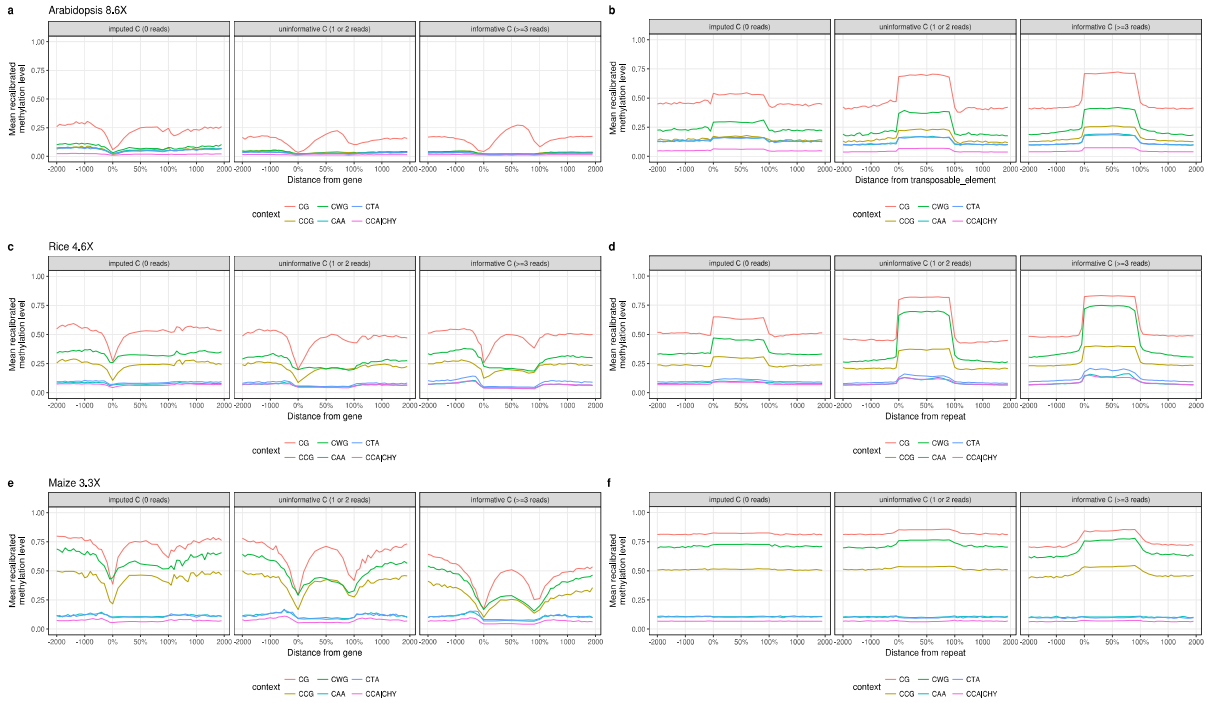


Figure 5-6 | Enrichment profiles for genes (left panels) and transposable elements or repeats (right panels). Sub-panels show the enrichment profiles for imputed (coverage = 0), uninformative (coverage = 1 or 2) and informative cytosines (coverage ≥ 3). See the Methods section for definition of the re-calibrated methylation level. (Source: Taudt et al. 2018, [138])

Saturation analysis for the performance assessment of imputed methylomes

`METHimpute` achieves high quality imputations by leveraging information from neighboring cytosines via the estimated distance-dependent transition probabilities (see Methods section). Therefore, confidence in the imputed calls is higher for cytosines that are closer to informative sites (SI-Figure 5-13). This spatial dependency remains high over distances of 10-40 bp and then decays to background levels. We reasoned that our imputation method may therefore be relatively robust even in shallow WGBS experiments, as long as enough measured cytosines are available to tag the methylation status of the underlying region.

To test this directly, we implemented a saturation analysis similar to Libertini et al. [155], where we compared high-coverage datasets with low-coverage subsets of these datasets. Bam files with mapped reads for the Arabidopsis, rice and maize replicates were merged to obtain samples with 23.2X, 18.6X and 7.2X coverage per cytosine per strand, respectively (SI-Table 1, online). These merged files were downsampled to generate a series of reduced datasets, ranging from 90% to 10% of the original data (SI-Table 3, online).

Upon downsampling, the proportion of cytosines with zero read coverage increased from 5% (23.2 X) to 31% (13.47M, 2.6 X) in *A. thaliana*, from 11% (18.6 X) to 40% (65.41M, 1.8 X) in *O. sativa* and from 14% (7.2 X) to 37% (52.07M, 2.2 X) in the *Z. mays* data (Figure 5-8|d-f). We ran `METHimpute` on each reduced dataset and calculated the F1-score in the status calls relative to those obtained with the full data. The F1-score is defined as the harmonic mean of precision and recall, and the status calls of the full dataset were assumed as ground truth.

Our analysis shows that performance remains remarkably high despite drastic decreases in sequencing depth (Figure 5-8|a-c, SI-Figure 5-14 with precision and recall, SI-Figure 5-15 F1-score per context). With data as low as 5 X coverage per cytosine (strand-specific), the F1-score was as high as 95% in Arabidopsis (U: 95%, M: 74%), 97% in rice (U: 97%, M: 88%) and 99% in maize (U: 99% M: 98%). In general, annotations with a large percentage of missing cytosines in the high coverage datasets were less accurately called upon downsampling (Figure 5-7). These include in particular transposable elements and repeats. The exception to this trend were 5' UTRs, which in all three species showed a large percentage of cytosines with missing data but a low amount of miscalled sites upon downsampling.

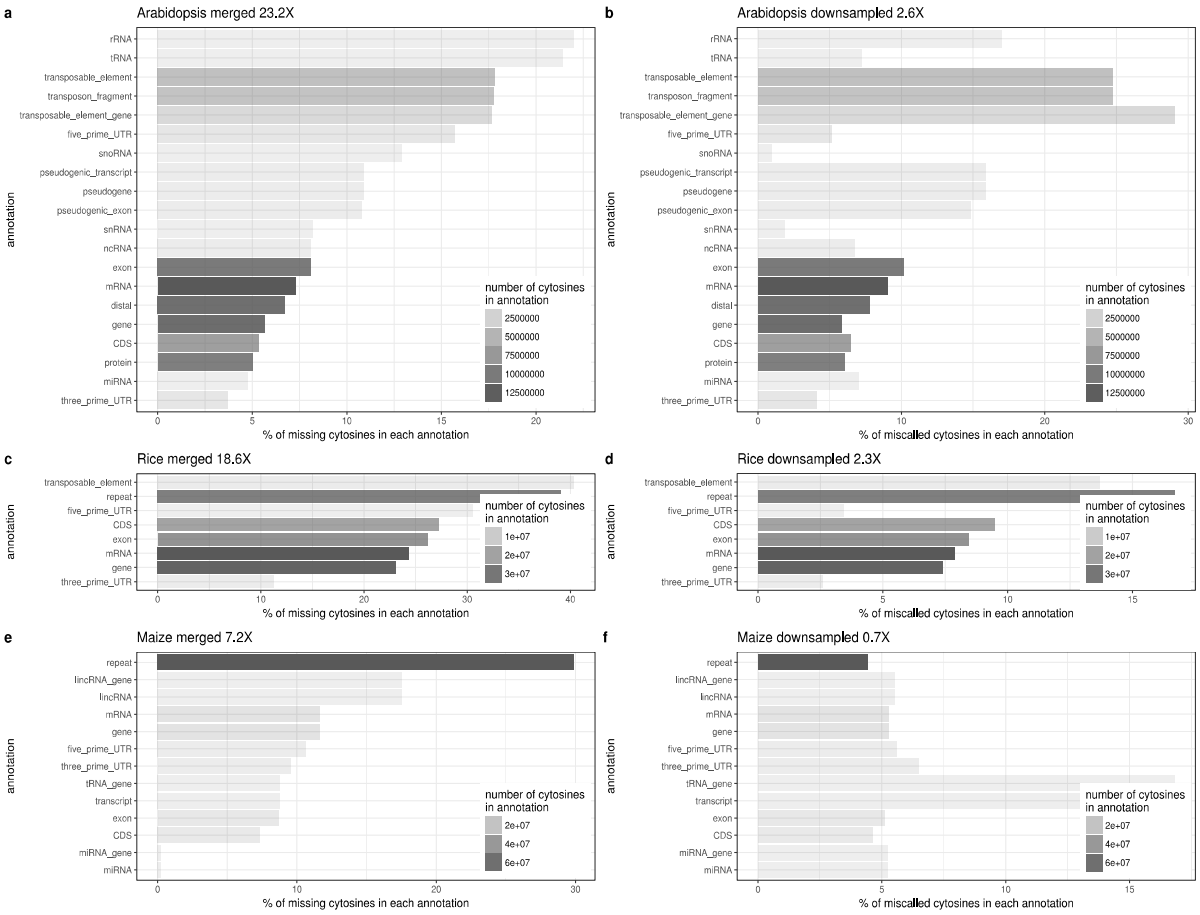


Figure 5-7 | Missing and miscalled cytosines per annotation. Left panels show the percentage of missing cytosines per annotation category, right panels show miscalled cytosines in a downsampled dataset. The transparency indicates the number of cytosines in the respective annotation. In all three species, transposable elements and repeats have a higher fraction of missing cytosines compared with genes. (Source: Taudt et al. 2018, [138])

To put the above accuracy analysis into perspective, we compared the HMM-based imputation method with a much simpler method based on the commonly used binomial test: Methylation states for informative cytosines (≥ 3 reads) were called with a binomial test, and methylation states for missing and uninformative cytosines (< 3 reads) were imputed by assigning the majority methylation state of covered cytosines of the same context in the 200 bp neighborhood of the missing or uninformative cytosine. Cytosines without any informative neighbors in a 200 bp neighborhood were not imputed and treated as “undefined”, and therefore counted as false negatives in the downsampled data if the full dataset was informative in these positions. We find that the accuracy obtained with this approach is less robust to average sequencing depth. With only 5 X data, the F1-score drops down to 93% (U: 93%, M: 74%) in Arabidopsis, 94% (U: 93%, M: 84%) in rice and 95% (U: 93%, M: 91%) in maize (Figure 5-8|a-c).

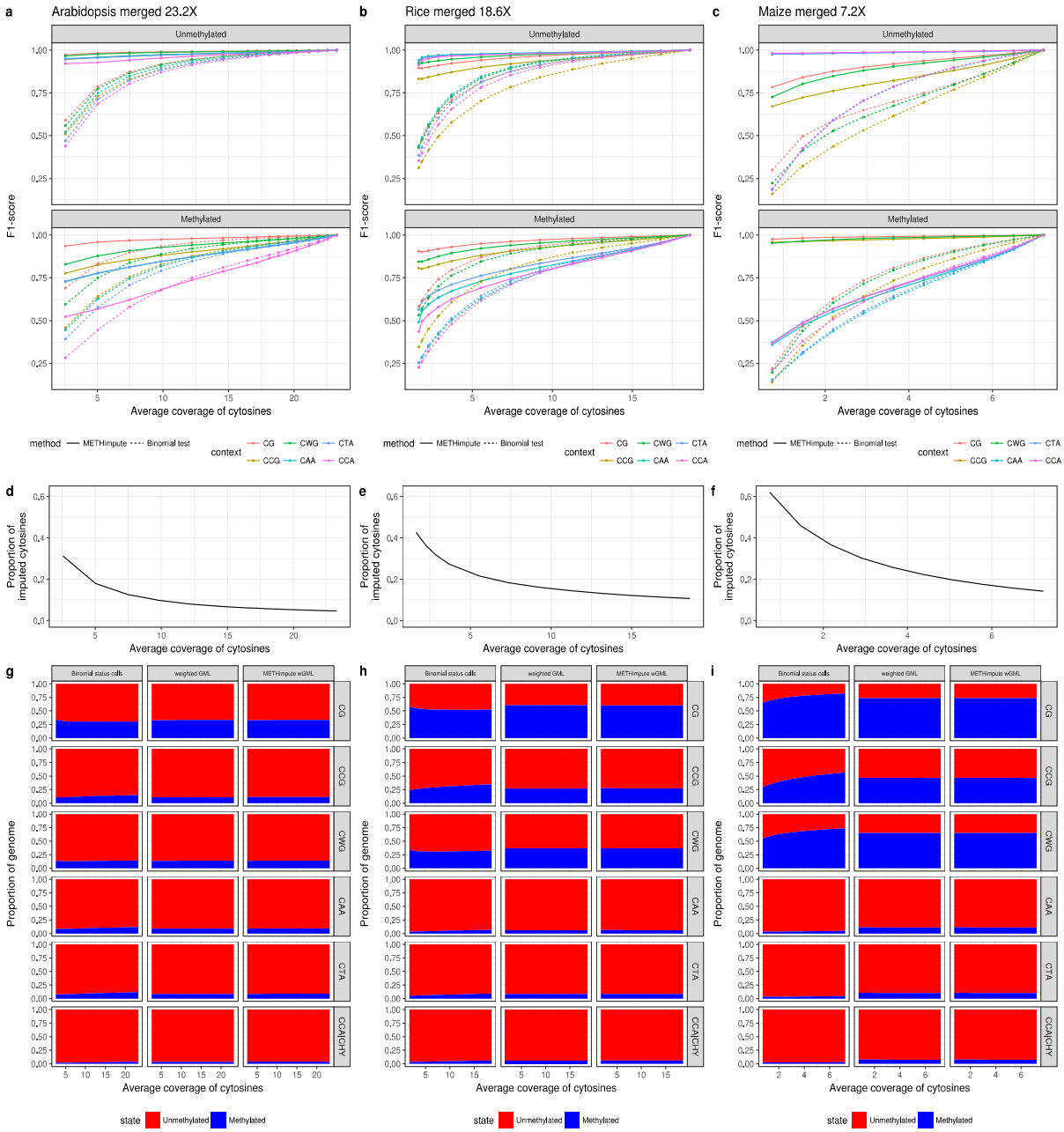


Figure 5-8 | Saturation analysis. **a-c** | F1-score for METHimpute and the binomial test, compared to the full sample, respectively. The F1-score is the harmonic mean of precision and recall. **d-f** | Proportion of imputed cytosines. **g-i** | Proportion of the genome in each state. The x-axes shows the average strand-specific coverage per cytosine. (Source: Taudt et al. 2018, [138])

Finally, we also considered the fidelity of the recalibrated methylation levels upon downsampling. Recalibrated methylation levels can be interpreted as the probability of observing a methylated read at a given position, and these recalibrated levels are highly correlated with original methylation levels: For Arabidopsis, rice and maize, the correlation (linear fit) was 0.91, 0.94 and 0.93, respectively (p-value ≤ 2e-16). To assess their fidelity upon downsampling, we calculated the correlation between recalibrated methylation levels per cytosine and per 100 bp window to the full coverage dataset, and compared that to the results obtained from the original methylation level (SI-Figure 5-16). Per-cytosine recalibrated methylation levels show slightly higher correlations than original methylation

levels, and with 10% of the original data the correlations for Arabidopsis, rice and maize are 0.89, 0.90 and 0.93, respectively. Window-based recalibrated methylation levels showed the same correlation performance as the original ones, with remarkably high correlations even when only 10% of the original data was retained (0.95, 0.95, 0.83 for Arabidopsis, rice, maize). These results suggest that recalibrated methylation levels can be used for downstream methylation analysis, since they are correlated to original methylation levels and are robust upon downsampling, while providing cytosine-level information even at low sequencing depth.

Overall, both for status calls and for recalibrated methylation levels, `METHimpute` produces robust results even at very low sequencing depth, suggesting that the algorithm offers a cost-effective solution for methylome studies of large genomes and for population-level studies involving a large number of samples.

Re-calibrated estimates of genome-wide and context-specific methylation levels

Plant species differ greatly in their genome-wide methylation levels (GMLs, *i.e.* the proportion of cytosines that are methylated) [100], [101]. In a recent survey of about 30 angiosperms, GMLs were found to be as low as 5% in *Theobroma cacao* to as high as 43% in *Beta vulgaris*, with a mean of about 16% [100], [135]. Much of this diversity appears to be the result of differences in genome size and repeat content, as well as differences in the efficiency of DNA methylation maintenance pathways [124]. Precise estimates of GMLs are important for studying the evolutionary forces that shape plant methylomes over short and long time-scales, and for understanding genome-epigenome co-evolution. However, obtaining GML estimates based on WGBS data is not trivial, as they are highly dependent on the method used for methylation status calling and on the depth of the sequencing experiment. In *A. thaliana*, for instance, reported GML estimates vary widely between studies. This dependency is even larger when considering context-specific GMLs (*i.e.* the proportion of methylated cytosines in a given context; CG-GMLs, CHG-GMLs, CHH-GMLs), with CHH-GMLs being by far the most variable between studies, with reported values ranging from as low as 1.51% [98] to as high as 3.91% [100].

In order to bypass many of the statistical issues in calling methylation states, especially in shallow WGBS data, recent studies have proposed so-called weighted genome-wide methylation levels (wGMLs) as a proxy for GMLs. A wGML is a non-statistical measure which is obtained by counting the number of methylated reads over the total number of reads at the genome-wide scale. Figure 5-8|g-i shows clearly that wGMLs are robust upon down-sampling in any sequence context in the *A. thaliana*, *O. sativa* and *Z. mays* data, thus justifying its use. By contrast, GMLs calculated from cytosine-level binomial status calls (*i.e.* # mC divided by all Cs) are highly unstable, particularly in non-CG contexts and when sequencing depth is low (Figure 5-8|g-i).

In order to assess whether the re-calibrated methylation levels provided by METHimpute can also be used to obtain robust estimates of GMLs, we calculated wGMLs by summing the per-cytosine re-calibrated methylation level genome-wide, weighted by coverage. Using this measure we find that METHimpute-derived wGMLs perform nearly identical to naive wGMLs, both in terms of robustness and magnitude (Figure 5-8|g-i, SI-Figure 5-17 with replicates). This demonstrates that METHimpute re-calibrated levels are consistent with original methylation levels and capture known biology not only at the individual cytosine level, but also aggregated over 100 bp windows and genome-wide, with the added advantage that they are available for all positions in the genome.

METHimpute facilitates insights into bisulfite conversion rates

One source of measurement noise in WGBS data is the bisulfite conversion procedure prior to sequencing. Bisulfite treatment of DNA is typically performed long enough so that all unmethylated cytosines are converted to uracils. The conversion success (or rate) is typically high. Most studies report conversion rates of about 0.99, implying that only about 1% of all unmethylated cytosines failed to convert. Knowledge of this rate is important not only to verify that bisulfite reaction was efficient but also to be able to separate biological signal from noise in downstream analyses of the data. Empirical estimates of the conversion rate are often obtained by including unmethylated chloroplast and virus genomes as controls in the WGBS workflow, and counting the number of non-converted cytosines from the mapped reads.

A helpful byproduct of the METHimpute fitting procedure is that the conversion rate can be directly estimated from the sequenced material without requiring auxiliary information from chloroplast or virus genomes. METHimpute achieves this in the HMM framework by estimating the probability, p_U, of finding a methylated read given that the underlying cytosine is unmethylated (see Methods, page 107), which can be used to derive the conversion rate. To obtain these rates we focus on estimates of p_U in context CG to exclude potential biases arising from the "fuzzy" maintenance of methylation at CHG and CHH sites. For *A. thaliana* and *Z. mays* our estimated conversion rates were 0.989 and 0.961, respectively, which is remarkably close to chloroplast-based estimates of 0.993 and 0.970.

Although bisulfite conversion kits and protocols have been optimized to achieve the highest conversion rate possible the specificity of the reaction is not perfect. A well-known trade-off is that some methylated cytosines can be accidentally converted to uracils, and are later falsely detected as unmethylated. Some controls (commercial or artificially methylated DNA fragments) are available to estimate this inappropriate conversion rate, but, to our knowledge, they are not systematically used in WGBS experiments. Some studies using such controls have shown that the inappropriate conversion rate (% of methylated cytosines converted to uracils) ranges from 0.09% to 6.1% depending on the kit and protocol used [156]–[158].

METHimpute approximates this value by estimating the parameter p_M for the M-component (see Methods, page 107), which can be used to calculate the probability of finding an unmethylated cytosine given that the underlying cytosine is truly methylated. Again, focusing on CG sites, we estimate the methylated cytosines conversion rate at 6.3%, 11.5% and 16% in *O. sativa*, *Z. mays* and *A. thaliana*, respectively. Although these estimates are close to the empirical rates reported in the literature, they are slightly biased upward most likely owing to the fact that the parameter p_M is partly confounded with methylation variation arising from cellular heterogeneity in the sampled tissues. We therefore suspect that our estimates become more accurate in situations where tissue heterogeneity is minimized.

Nonetheless, the ability of METHimpute to provide an accurate estimate of the conversion rate for unmethylated cytosines and an upper-bound estimate for methylated cytosines could be utilized to calibrate WGBS experiments in the laboratory when no controls are available.

Discussion

A key advantage of WGBS over alternative measurement technologies is its ability to provide cytosine-level measurements from bulk and – more recently – also from single-cell data. Since its first application in the model plant *A. thaliana* in 2008 [153], [154], WGBS has become an integral tool for studying the methylomes of increasingly large plant genomes and for surveying patterns of natural methylome variation within and among plant species. However, the relatively high costs associated with this technology pose limits on the sequencing depths that can be achieved within most experimental budgets. A typical solution is to sequence methylomes far below saturation, which results in substantial measurement noise and missing data at the level of individual cytosines.

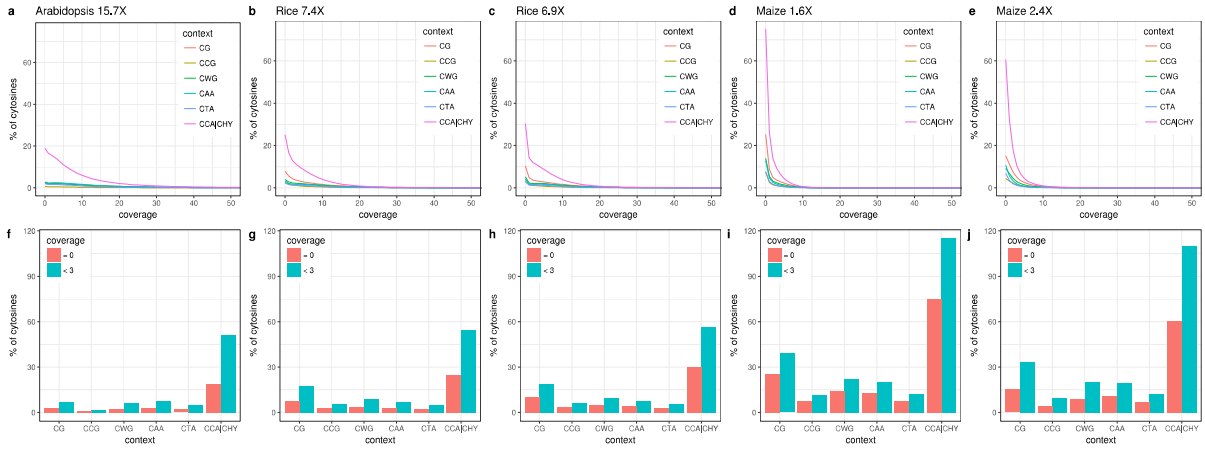
Here we introduced METHimpute, an imputation-based HMM for the reconstruction of complete methylomes from shallow or deep WGBS data. Our analyses show that the algorithm can impute the methylation status of cytosines with missing data (*i.e.* zero read coverage) or uninformative coverage (*i.e.* coverage of less than 3 reads), as well as their re-calibrated methylation levels. We demonstrate that these imputations are not only statistically robust, but also biologically meaningful. Our estimates suggest that routine use of this algorithm could reduce sequencing costs of typically sized methylome experiments without a substantial loss of biological information. The method works with small, streamlined genomes like that of Arabidopsis but also with large, repeat-rich genomes like those of most commercial crops, thus making it a flexible software tool for the analysis of DNA methylomes of a wide spectrum of species.

We show that our approach imputes high-confidence methylation calls for cytosines that are sufficiently close to informative cytosines (< 40 bp). For cytosines in widely uncovered regions, our approach imputes low-confidence calls which might be filtered out in downstream analyses, since they do not contain more information than background frequencies of methylation. A threshold for filtering can be determined from the asymptotic behavior of the maximum posterior probability as in SI-Figure 5-13.

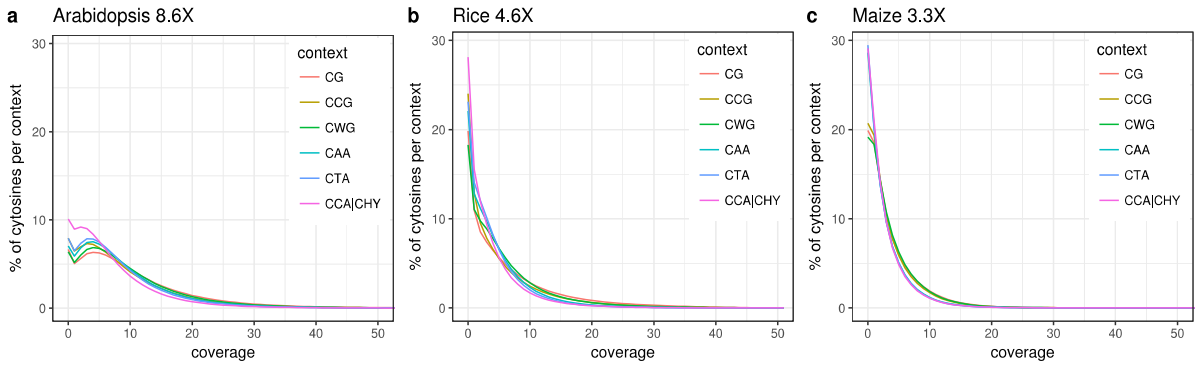
We recommend the use of `METHimpute` instead of the binomial test for the analysis of WGBS data whenever methylation status calls are required. Furthermore, `METHimpute` solves the problem of missing data in populatio n epigenetic studies, which will facilitate the estimation of epigenetic mutation rates and methylation site frequency spectrum analyses. `METHimpute` is implemented as an R-package and seamlessly integrates with the extensive bioinformatic tool sets available through Bioconductor. The algorithm has been extensively tested in plants, but it should also be applicable in non-plant species.

Supplemental Material

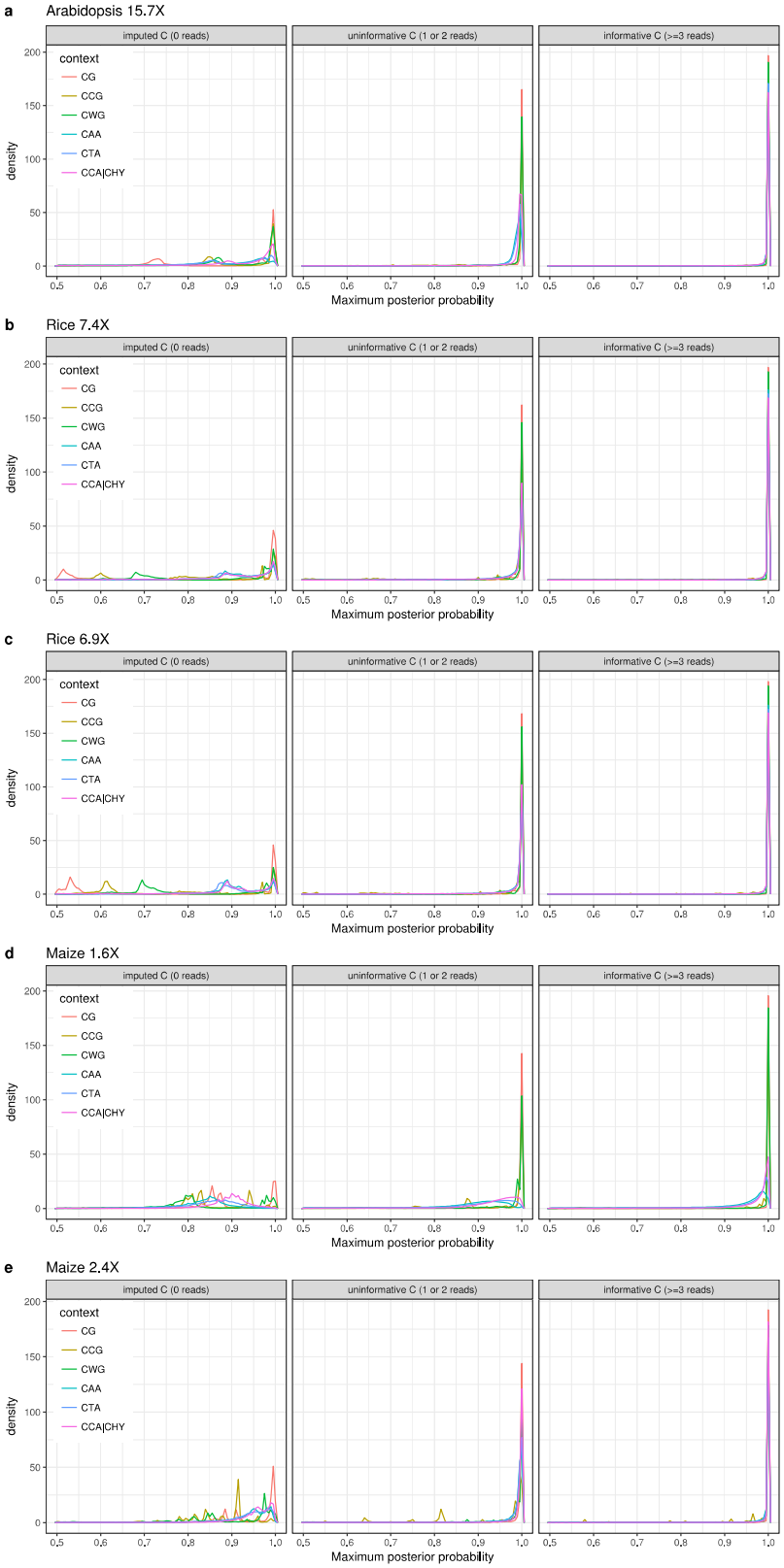
Supplemental Figures



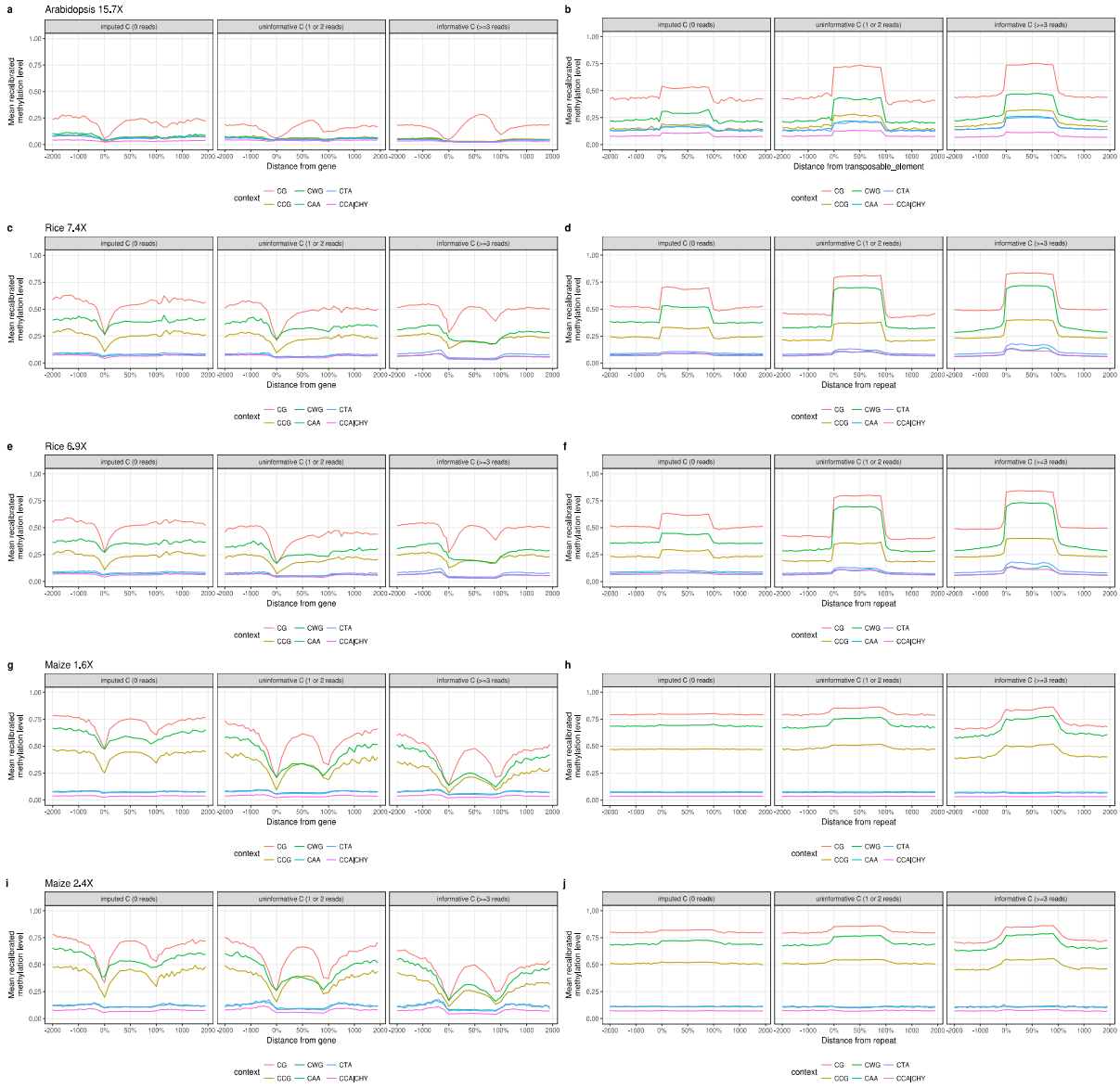
SI-Figure 5-9 | Coverage distributions for replicates. **a-e** | Percentage of cytosines with X coverage (strand-specific). **f-j** | Percentage of cytosines with missing data (red) and "uninformative" coverage (green), defined as less than three reads. (Source: Taudt et al. 2018, [138])



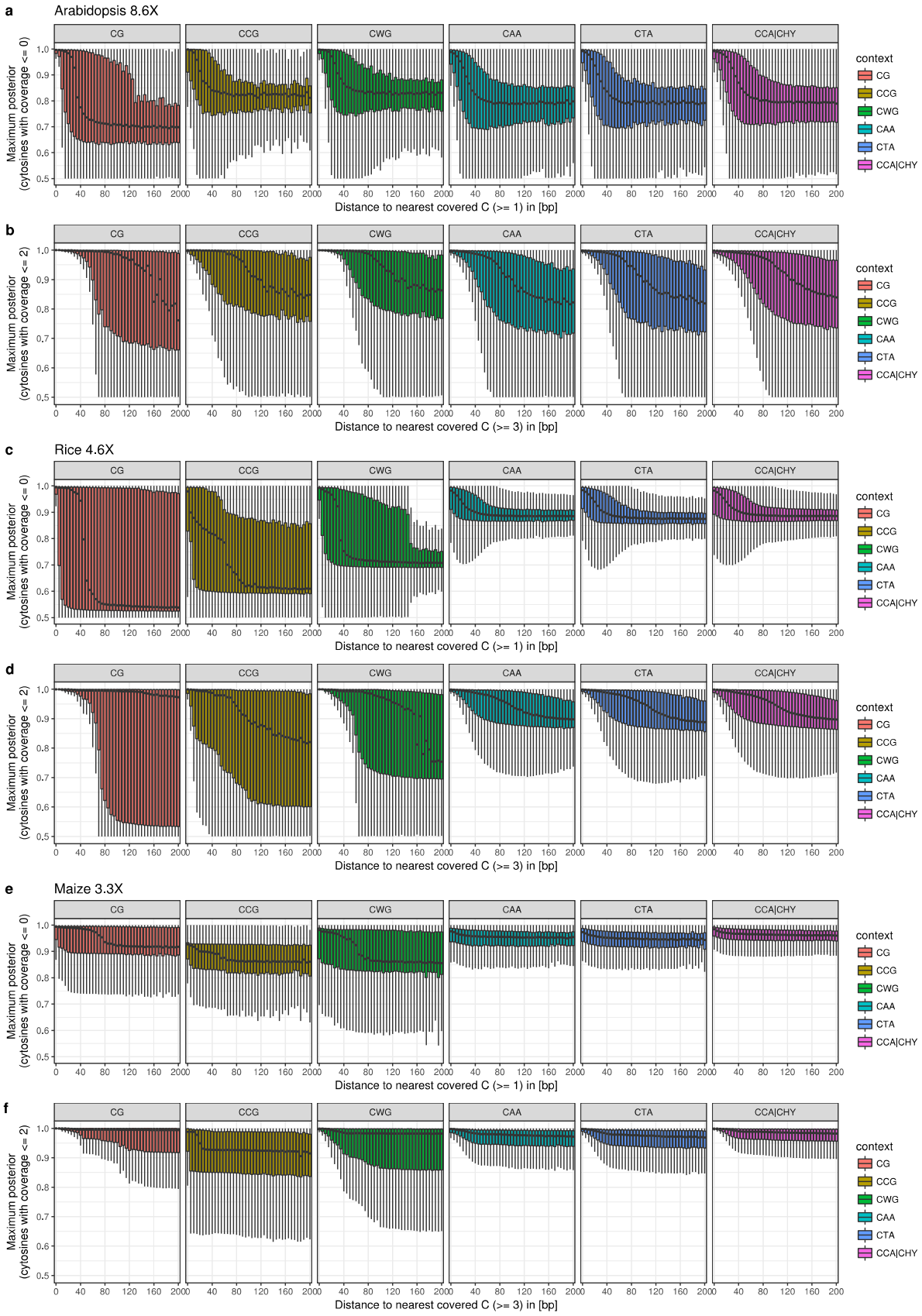
SI-Figure 5-10 | Context-specific coverage distributions. Percentage of cytosines with X coverage (strand-specific), normalized by context. The CHH context has a higher percentage of missing and uninformative cytosines. (Source: Taudt et al. 2018, [138])



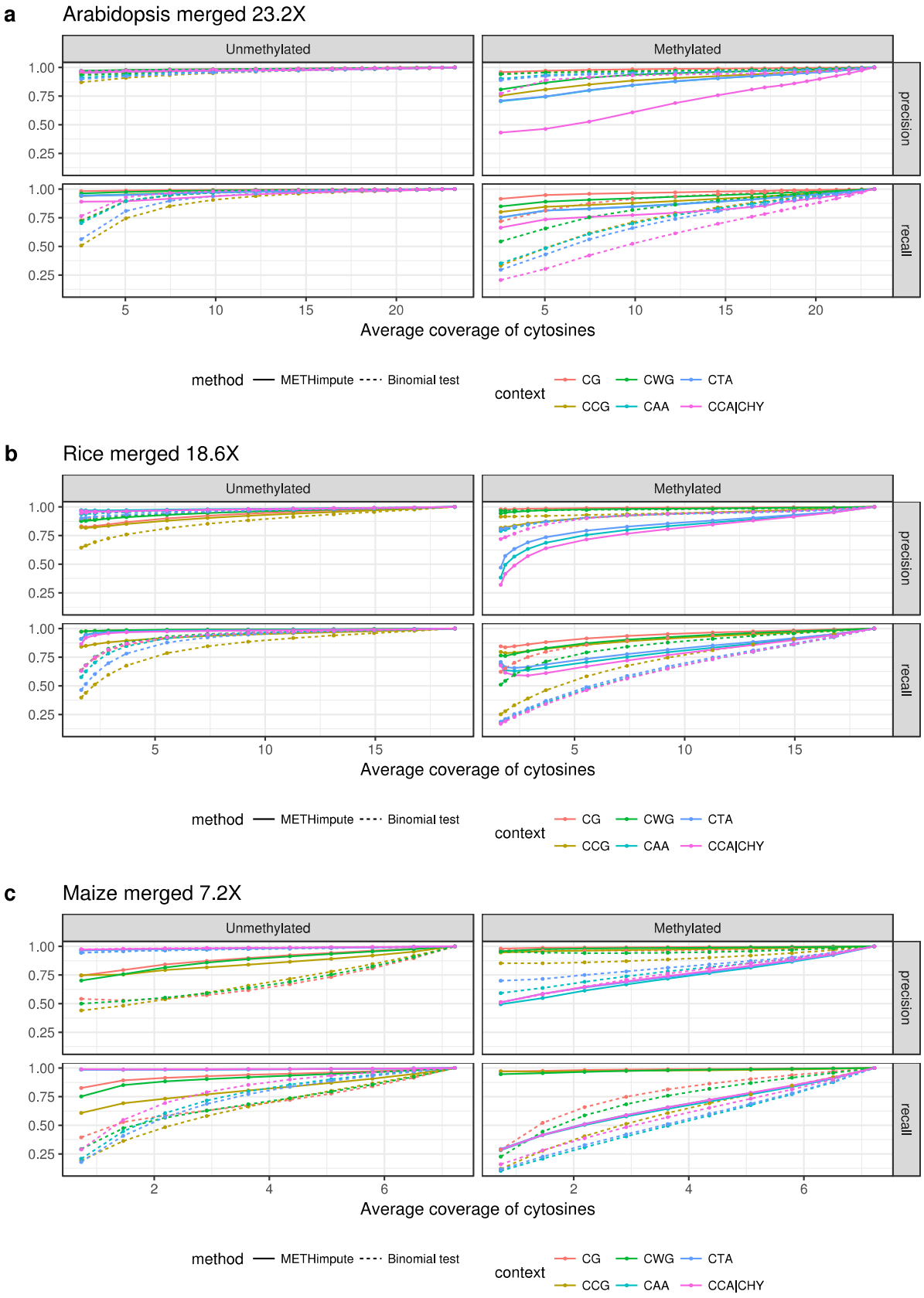
SI-Figure 5-11 | Maximum posterior distributions for replicates for imputed cytosines (coverage = 0), uninformative cytosines (coverage = 1 or 2) and informative cytosines (coverage >= 3). The maximum posterior probability, *i.e.* the confidence in the methylation status calls, is generally lower for sites with less coverage. (Source: Taudt et al. 2018, [138])



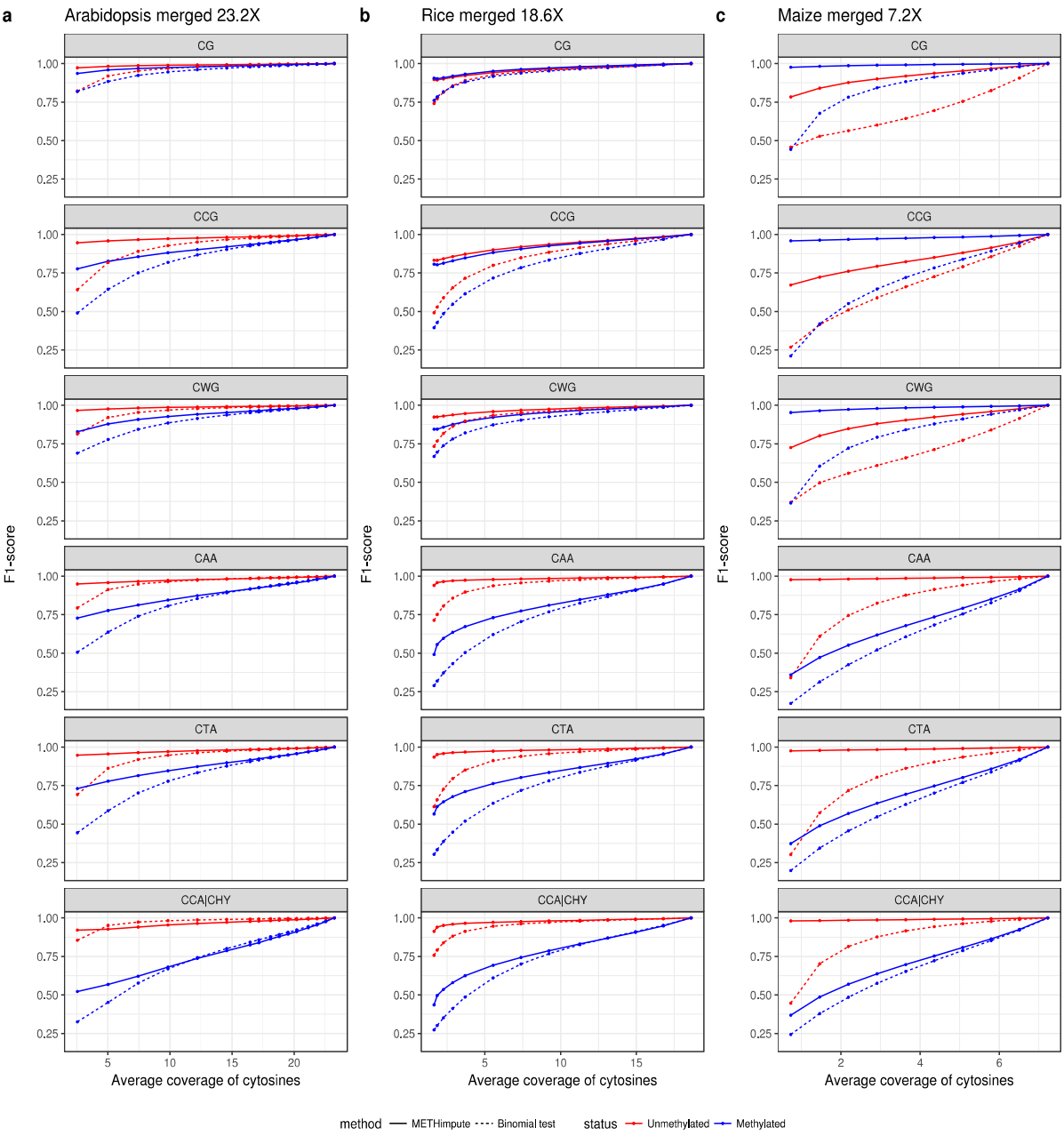
SI-Figure 5-12 | Enrichment profiles for replicates for genes (left panels) and transposable elements or repeats (right panels). Sub-panels show the enrichment profiles for imputed (coverage = 0), uninformative (coverage = 1 or 2) and informative cytosines (coverage >= 3). (Source: Taudt et al. 2018, [138])



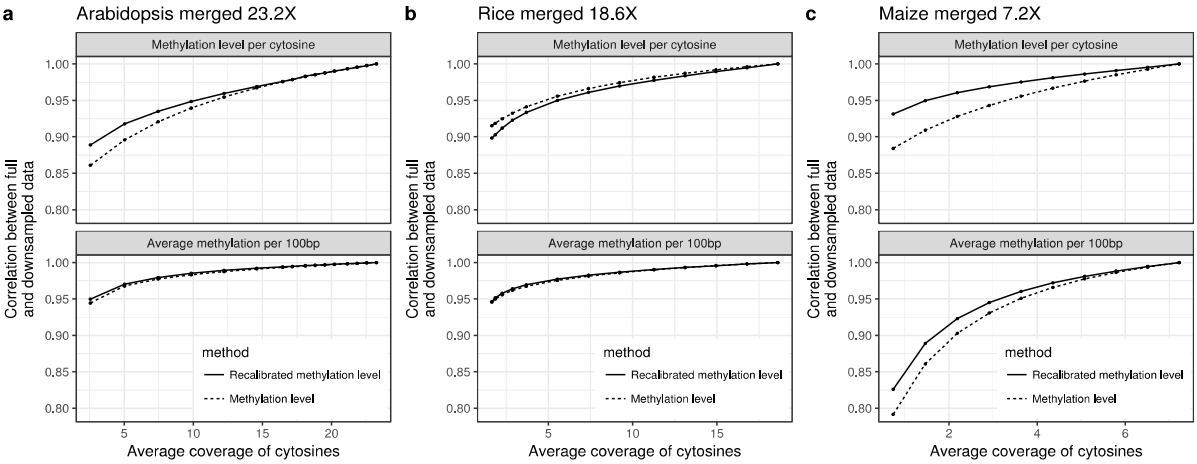
SI-Figure 5-13 | Maximum posterior vs. distance. The maximum posterior probability (y-axis) is plotted against the distance to the nearest covered cytosine (x-axis). We observe that the maximum posterior probability, *i.e.* the confidence in the methylation status calls, decays to background levels if the nearest covered cytosine is more than 40-80 bp away. (Source: Taudt et al. 2018, [138])



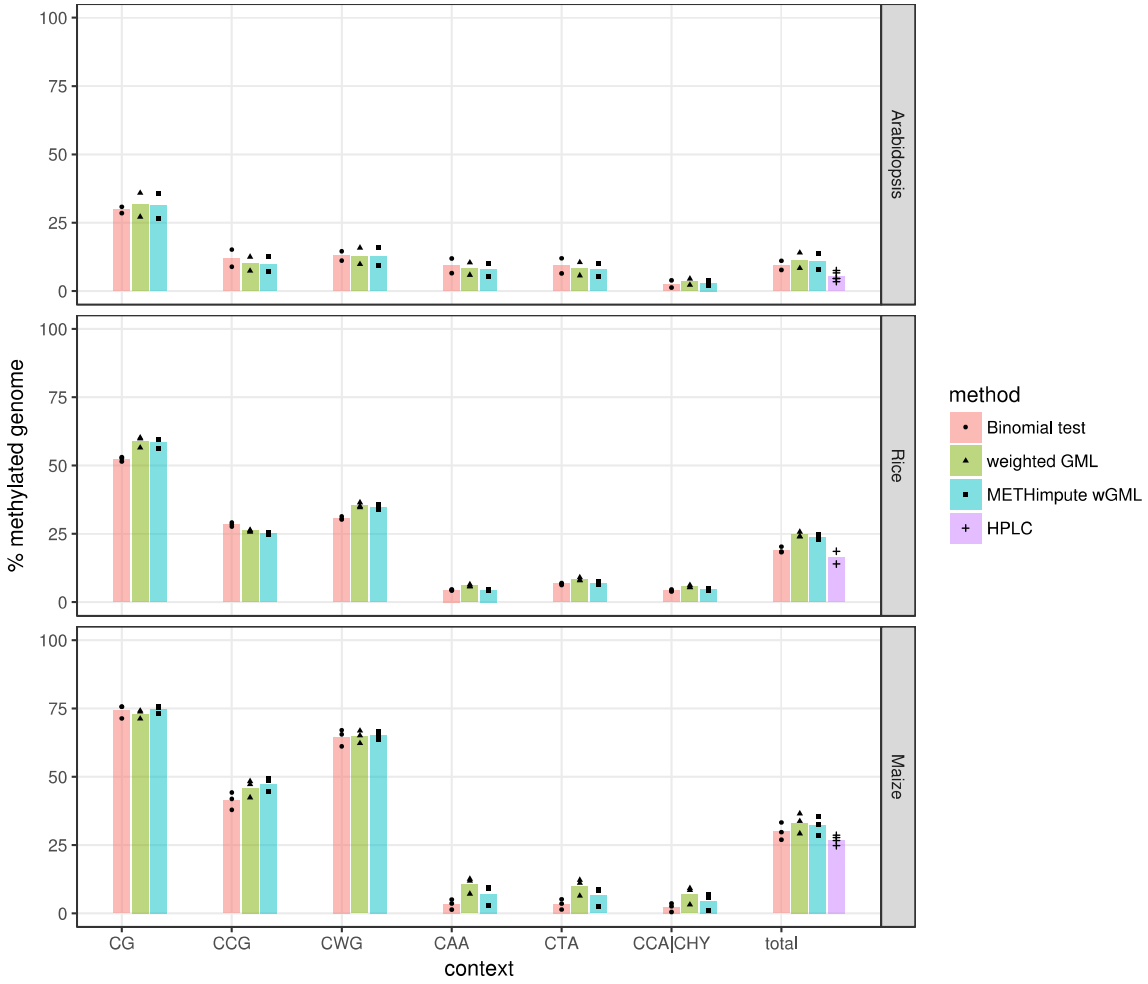
SI-Figure 5-14 | Saturation analysis. Precision and recall for the methylation status calls for `METHimpute` and the binomial test, compared to the full sample, respectively. **a** | Arabidopsis, **b** | Rice, **c** | Maize. (Source: Taudt et al. 2018, [138])



SI-Figure 5-15 | Saturation analysis. F1-score for the methylation status calls for `METHimpute` and the binomial test, compared to the full sample, respectively. F1-score is shown for **a** | Arabidopsis, **b** | Rice, **c** | Maize. Sub-panels show the different contexts. (Source: Taudt et al. 2018, [138])



SI-Figure 5-16 | Saturation analysis. Correlation between the full and downsampled datasets for original methylation levels and `METHimpute` recalibrated methylation levels. The correlation is shown for **a** | Arabidopsis, **b** | Rice, **c** | Maize. Top-panels show correlations for individual cytosines, bottom-panels show the correlation for levels averaged (weighted by coverage) over 100 bp windows. (Source: Taudt et al. 2018, [138])



SI-Figure 5-17 | Comparison of GMLs. Genome-wide methylation levels (GMLs) calculated by different methods in Arabidopsis, rice and maize. The bar chart indicates the mean methylation level among replicates, dots indicate the methylation level for individual replicates. Differences in GML between HPLC and the other methods are not significant (t-test, $p > 0.05$). (Source: Taudt et al. 2018, [138])

Bibliography

[1] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, “A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains,” *Ann. Math. Stat.*, vol. 41, no. 1, pp. 164–171, Feb. 1970.

[2] G. McLachlan and D. Peel, *Finite Mixture Models*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2000.

[3] M. Kircher and J. Kelso, “High-throughput DNA sequencing - concepts and limitations,” *BioEssays*, vol. 32, no. 6, pp. 524–536, May 2010.

[4] H. P. J. Buermans and J. T. den Dunnen, “Next generation sequencing technology: Advances and applications,” *Biochim. Biophys. Acta - Mol. Basis Dis.*, vol. 1842, no. 10, pp. 1932–1941, Oct. 2014.

[5] S. Yohe and B. Thyagarajan, “Review of Clinical Next-Generation Sequencing,” *Arch. Pathol. Lab. Med.*, vol. 141, no. November, p. arpa.2016-0501-RA, Nov. 2017.

[6] L. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proc. IEEE*, 1989.

[7] B. Bakker, H. van den Bos, P. M. Lansdorp, and F. Foijer, “How to count chromosomes in a cell: An overview of current and novel technologies,” *BioEssays*, p. n/a-n/a, 2015.

[8] B. Bakker *et al.*, “Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies,” *Genome Biol.*, vol. 17, no. 1, p. 115, May 2016.

[9] H. van den Bos *et al.*, “Single-cell whole genome sequencing reveals no evidence for common aneuploidy in normal and Alzheimer’s disease neurons,” *Genome Biol.*, vol. 17, no. 1, p. 116, May 2016.

[10] S. M. Teo, Y. Pawitan, C. S. Ku, K. S. Chia, and A. Salim, “Statistical challenges associated with detecting copy number variations with next-generation sequencing.,” *Bioinformatics*, vol. 28, no. 21, pp. 2711–8, Nov. 2012.

[11] M. Lawrence *et al.*, “Software for computing and annotating genomic ranges.,” *PLoS Comput. Biol.*, vol. 9, no. 8, p. e1003118, Jan. 2013.

[12] T. Daley and A. D. Smith, “Predicting the molecular complexity of sequencing libraries.,” *Nat. Methods*, vol. 10, no. 4, pp. 325–7, 2013.

[13] C. Fraley and A. E. Raftery, “Model-based Clustering, Discriminant Analysis and Density Estimation,” *J. Am. Stat. Assoc.*, vol. 97, pp. 611–631, 2002.

[14] C. Fraley, A. E. Raftery, T. B. Murphy, and L. Scrucca, “mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation,” no. 597. 2012.

[15] T. Garvin *et al.*, “Interactive analysis and quality assessment of single-cell copy-number variations,” *Nat. Methods*, vol. 12, no. 11, pp. 1058–1060, Sep. 2015.

[16] N. Navin *et al.*, “Tumour evolution inferred by single-cell sequencing,” *Nature*, vol. 472, no. 7341, pp. 90–94, Apr. 2011.

[17] E. Falconer and P. M. Lansdorp, “Strand-seq: a unifying tool for studies of chromosome segregation.,” *Semin. Cell Dev. Biol.*, vol. 24, no. 8–9, pp. 643–52, 2013.

[18] E. Falconer *et al.*, “DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution,” *Nat. Methods*, vol. 9, no. 11, pp. 1107–1112, Oct. 2012.

[19] M. Hills, K. O’Neill, E. Falconer, R. Brinkman, and P. M. Lansdorp, “BAIT: Organizing genomes and mapping rearrangements in single cells.,” *Genome Med.*, vol. 5, no. 9, p. 82, Jan. 2013.

[20] A. D. Sanders, M. Hills, D. Porubský, V. Guryev, E. Falconer, and P. M. Lansdorp, “Characterizing polymorphic inversions in human genomes by single-cell sequencing.,” *Genome Res.*, vol. 26, no. 11, pp. 1575–1587, Nov. 2016.

[21] D. Porubsky, “Haplotype resolved genomes: Computational challenges and applications,” University of Groningen, 2017.

[22] N. James and D. Matteson, “ecp: An R Package for nonparametric multiple change point analysis of multivariate data,” *J. Stat. Softw.*, vol. 62, no. 1, pp. 1–25, Jan. 2015.

[23] P. J. Park, “ChIP-seq: advantages and challenges of a maturing technology.,” *Nat. Rev. Genet.*, vol. 10, no. 10, pp. 669–80, Oct. 2009.

[24] P. W. Laird, “Principles and challenges of genomewide DNA methylation analysis.,” *Nat. Rev. Genet.*, vol. 11, no. 3, pp. 191–203, Mar. 2010.

[25] D. Adams *et al.*, “BLUEPRINT to decode the epigenetic signature written in blood,” *Nat. Biotechnol.*, vol. 30, no. 3, pp. 224–226, Mar. 2012.

[26] B. E. Bernstein *et al.*, “The NIH Roadmap Epigenomics Mapping Consortium.,” *Nat. Biotechnol.*, vol. 28, no. 10, pp. 1045–8, Oct. 2010.

[27] R. E. Consortium *et al.*, “Integrative analysis of 111 reference human epigenomes,” *Nature*, vol. 518, pp. 317–330, 2015.

[28] J. Ernst and M. Kellis, “Discovery and characterization of chromatin states for systematic annotation of the human genome.,” *Nat. Biotechnol.*, vol. 28, no. 8, pp. 817–25, Aug. 2010.

[29] M. Kasowski *et al.*, “Extensive variation in chromatin states across humans.,” *Science*, vol. 342, no. 6159, pp. 750–2, Nov. 2013.

[30] V. W. Zhou, A. Goren, and B. E. Bernstein, “Charting histone modifications and the functional organization of mammalian genomes.,” *Nat. Rev. Genet.*, vol. 12, no. 1, pp. 7–18, Jan. 2011.

[31] T. Kouzarides, “Chromatin modifications and their function.,” *Cell*, vol. 128, no. 4, pp. 693–705, Feb. 2007.

[32] B. E. Bernstein, A. Meissner, and E. S. Lander, “The mammalian epigenome.,” *Cell*, vol. 128, no. 4, pp. 669–81, Feb. 2007.

[33] B. D. Strahl and C. D. Allis, “The language of covalent histone modifications.,” *Nature*, vol. 403, no. 6765, pp. 41–5, Jan. 2000.

[34] T. Jenuwein and C. D. Allis, “Translating the histone code.,” *Science*, vol. 293, no. 5532, pp. 1074–80, Aug. 2001.

[35] X. Li *et al.*, “High-resolution mapping of epigenetic modifications of the rice genome uncovers interplay between DNA methylation, histone methylation, and gene expression,” *Plant Cell*, vol. 20, no. 2, pp. 259–276, 2008.

[36] G. Hon, B. Ren, and W. Wang, “ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome.,” *PLoS Comput. Biol.*, vol. 4, no. 10, p. e1000201, Oct. 2008.

[37] Z. Wang *et al.*, “Combinatorial patterns of histone acetylations and methylations in the human genome.,” *Nat. Genet.*, vol. 40, no. 7, pp. 897–903, Jul. 2008.

[38] G. Hon, W. Wang, and B. Ren, “Discovery and annotation of functional chromatin signatures in the human genome.,” *PLoS Comput. Biol.*, vol. 5, no. 11, p. e1000566, Nov. 2009.

[39] S. Roy *et al.*, “Identification of functional elements and regulatory circuits by Drosophila modENCODE.,” *Science*, vol. 330, no. 6012, pp. 1787–97, Dec. 2010.

[40] P. V Kharchenko *et al.*, “Comprehensive analysis of the chromatin landscape in Drosophila melanogaster.,” *Nature*, vol. 471, no. 7339, pp. 480–5, Mar. 2011.

[41] N. C. Riddle *et al.*, “Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin,” *Genome Res.*, vol. 21, no. 2, pp. 147–163, Dec. 2010.

[42] M. B. Gerstein *et al.*, “Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project.,” *Science*, vol. 330, no. 6012, pp. 1775–87, Dec. 2010.

[43] G. J. Filion *et al.*, “Systematic protein location mapping reveals five principal chromatin types in Drosophila cells.,” *Cell*, vol. 143, no. 2, pp. 212–24, Oct. 2010.

[44] F. Roudier *et al.*, “Integrative epigenomic mapping defines four main chromatin states in Arabidopsis,” *EMBO J.*, vol. 30, no. 10, pp. 1928–1938, May 2011.

[45] T. Liu *et al.*, “Broad chromosomal domains of histone modification patterns in C. elegans.,” *Genome Res.*, vol. 21, no. 2, pp. 227–36, Feb. 2011.

[46] M. M. Hoffman, O. J. Buske, J. Wang, Z. Weng, J. A. Bilmes, and W. S. Noble, “Unsupervised pattern discovery in human chromatin structure through genomic segmentation.,” *Nat. Methods*, vol. 9, no. 5, pp. 473–6, May 2012.

[47] W. K. M. Lai and M. J. Buck, “An integrative approach to understanding the combinatorial histone code at functional elements.,” *Bioinformatics*, vol. 29, no. 18, pp. 2231–7, Sep. 2013.

[48] C. Luo, D. J. Sidote, Y. Zhang, R. A. Kerstetter, T. P. Michael, and E. Lam, “Integrative analysis of chromatin states in Arabidopsis identified potential regulatory mechanisms for natural antisense transcript production.,” *Plant J.*, vol. 73, no. 1, pp. 77–90, Jan. 2013.

[49] J. Sequeira-Mendes *et al.*, “The Functional Topography of the Arabidopsis Genome Is Organized in a Reduced Number of Linear Motifs of Chromatin States.,” *Plant Cell*, vol. 26, no. 6, pp. 2351–2366, Jun. 2014.

[50] K. Baker *et al.*, “Chromatin state analysis of the barley epigenome reveals a higher-order structure defined by H3K27me1 and H3K27me3 abundance,” *Plant J.*, vol. 84, no. 1, pp. 111–124, Oct. 2015.

[51] A. Taudt, M. A. Nguyen, M. Heinig, F. Johannes, and M. Colome-Tatche, “chromstaR: Tracking combinatorial chromatin state dynamics in space and time,” Cold Spring Harbor Labs Journals, Feb. 2016.

[52] D. Lara-Astiaso *et al.*, “Chromatin state dynamics during blood formation.,” *Science*, vol. 55, no. 233348, pp. 1–10, 2014.

[53] D. Leung *et al.*, “Integrative analysis of haplotype-resolved epigenomes across human tissues,” *Nature*, vol. 518, pp. 350–354, 2015.

[54] R. Andersson, “Promoter or enhancer, what’s the difference? Deconstruction of established distinctions and presentation of a unifying model.,” *Bioessays*, vol. 37, no. 3, pp. 314–23, Mar. 2015.

[55] X. Zeng, R. Sanalkumar, E. H. Bresnick, H. Li, Q. Chang, and S. Keleş, “jMOSAiCS: joint analysis of multiple ChIP-seq datasets.,” *Genome Biol.*, vol. 14, no. 4, p. R38, Jan. 2013.

[56] K.-J. Won *et al.*, “Comparative annotation of functional regions in the human genome using epigenomic data.,” *Nucleic Acids Res.*, vol. 41, no. 8, pp. 4423–32, Apr. 2013.

[57] A. Taudt, M. Colomé-Tatché, and F. Johannes, “Genetic sources of population epigenomic variation,” *Nat. Rev. Genet.*, vol. 17, no. 6, pp. 319–332, May 2016.

[58] T. S. Mikkelsen *et al.*, “Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.,” *Nature*, vol. 448, no. 7153, pp. 553–60, Aug. 2007.

[59] A. Barski *et al.*, “High-resolution profiling of histone methylations in the human genome.,” *Cell*, vol. 129, no. 4, pp. 823–37, May 2007.

[60] C. M. Koch *et al.*, “The landscape of histone modifications across 1% of the human genome in five human cell lines.,” *Genome Res.*, vol. 17, no. 6, pp. 691–707, Jun. 2007.

[61] A. Huda, L. Mariño-Ramírez, and I. K. Jordan, “Epigenetic histone modifications of human transposable elements: genome defense versus exaptation.,” *Mob. DNA*, vol. 1, no. 1, p. 2, Jan. 2010.

[62] G. Robertson *et al.*, “Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.,” *Nat. Methods*, vol. 4, no. 8, pp. 651–7, Aug. 2007.

[63] D. K. Pokholok *et al.*, “Genome-wide map of nucleosome acetylation and methylation in yeast.,” *Cell*, vol. 122, no. 4, pp. 517–27, Aug. 2005.

[64] C. Rintisch *et al.*, “Natural variation of histone modification and its impact on gene expression in the rat genome.,” *Genome Res.*, vol. 24, no. 6, pp. 942–53, Jun. 2014.

[65] B. E. Bernstein, E. Birney, I. Dunham, E. D. Green, C. Gunter, and M. Snyder, “An integrated encyclopedia of DNA elements in the human genome.,” *Nature*, vol. 489, no. 7414, pp. 57–74, Sep. 2012.

[66] M. M. Hoffman *et al.*, “Integrative annotation of chromatin elements from ENCODE data.,” *Nucleic Acids Res.*, vol. 41, no. 2, pp. 827–41, Jan. 2013.

[67] J. Ernst and M. Kellis, “ChromHMM: automating chromatin-state discovery and characterization.,” *Nat. Methods*, vol. 9, no. 3, pp. 215–216, Mar. 2012.

[68] J. Biesinger, Y. Wang, and X. Xie, “Discovering and mapping chromatin states using a tree hidden Markov model.,” *BMC Bioinformatics*, vol. 14 Suppl 5, p. S4, Jan. 2013.

[69] K.-A. Sohn, J. W. K. Ho, D. Djordjevic, H.-H. Jeong, P. J. Park, and J. H. Kim, “hiHMM: Bayesian non-parametric joint inference of chromatin state maps.,” *Bioinformatics*, p. btv117-, Feb. 2015.

[70] J. Song and K. C. Chen, “Spectacle: fast chromatin state annotation using spectral learning.,” *Genome Biol.*, vol. 16, no. 1, p. 33, 2015.

[71] A. Mammana and H.-R. Chung, “Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome.,” *Genome Biol.*, vol. 16, no. 1, p. 151, Jan. 2015.

[72] A. Yen and M. Kellis, “Systematic chromatin state comparison of epigenomes associated with diverse properties including sex and tissue type,” *Nat. Commun.*, vol. 6, p. 7973, 2015.

[73] H. Ji, X. Li, Q. -f. Wang, and Y. Ning, “Differential principal component analysis of ChIP-seq,” *Proc. Natl. Acad. Sci.*, vol. 110, no. 17, pp. 6789–6794, Apr. 2013.

[74] N. U. Rashid, P. G. Giresi, J. G. Ibrahim, W. Sun, and J. D. Lieb, “ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions.,” *Genome Biol.*, vol. 12, no. 7, p. R67, Jan. 2011.

[75] C. Spyrou, R. Stark, A. G. Lynch, and S. Tavaré, “BayesPeak: Bayesian analysis of ChIP-seq data.,” *BMC Bioinformatics*, vol. 10, p. 299, Jan. 2009.

[76] A. van der Graaf *et al.*, “Rate, spectrum, and evolutionary dynamics of spontaneous epimutations.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 112, no. 21, pp. 6676–81, May 2015.

[77] M. Sklar, *Fonctions de répartition à n dimensions et leurs marges*. 1959.

[78] M. Heinig *et al.*, “histoneHMM: Differential analysis of histone modifications with broad genomic footprints,” *BMC Bioinformatics*, vol. 16, no. 1, p. 60, Feb. 2015.

[79] B. Renard and M. Lang, “Use of a Gaussian copula for multivariate extreme value analysis: Some case studies in hydrology,” *Adv. Water Resour.*, vol. 30, no. 4, pp. 897–912, 2007.

[80] B. L. Kidder, G. Hu, and K. Zhao, “ChIP-Seq: Technical Considerations for Obtaining High Quality Data,” *Nat. Immunol.*, vol. 12, no. 10, pp. 918–922, Oct. 2013.

[81] Y. Zhang *et al.*, “Model-based analysis of ChIP-Seq (MACS).,” *Genome Biol.*, vol. 9, no. 9, p. R137, Jan. 2008.

[82] C. Zang, D. E. Schones, C. Zeng, K. Cui, K. Zhao, and W. Peng, “A clustering approach for identification of enriched domains from histone modification ChIP-Seq data.,” *Bioinformatics*, vol. 25, no. 15, pp. 1952–8, Aug. 2009.

[83] H. Xing *et al.*, “Genome-Wide Localization of Protein-DNA Binding and Histone Modification by a Bayesian Change-Point Method with ChIP-seq Data,” *PLoS Comput. Biol.*, vol. 8, no. 7, p. e1002613, Jul. 2012.

[84] A. Harmanci, J. Rozowsky, and M. Gerstein, “MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework.,” *Genome Biol.*, vol. 15, no. 474, pp. 1–15, 2014.

[85] M. Micsinai, F. Parisi, F. Strino, P. Asp, B. D. Dynlacht, and Y. Kluger, “Picking ChIP-seq peak detectors for analyzing chromatin modification experiments.,” *Nucleic Acids Res.*, vol. 40, no. 9, p. e70, May 2012.

[86] X. Robin *et al.*, “pROC: an open-source package for R and S+ to analyze and compare ROC curves,” *BMC Bioinformatics*, vol. 12, no. 1, p. 77, 2011.

[87] B. E. Bernstein *et al.*, “Genomic maps and comparative analysis of histone modifications in human and mouse.,” *Cell*, vol. 120, no. 2, pp. 169–81, Jan. 2005.

[88] N. D. Heintzman *et al.*, “Histone modifications at human enhancers reflect global cell-type-specific gene expression.,” *Nature*, vol. 459, no. 7243, pp. 108–12, May 2009.

[89] R. Andersson *et al.*, “An atlas of active enhancers across human cell types and tissues.,” *Nature*, vol. 507, no. 7493, pp. 455–61, 2014.

[90] J. R. Dixon *et al.*, “Chromatin architecture reorganization during stem cell differentiation,” *Nature*, vol. 518, no. 7539, pp. 331–336, 2015.

[91] V. Amin *et al.*, “Epigenomic footprints across 111 reference epigenomes reveal tissue-specific epigenetic regulation of lincRNAs,” *Nat. Commun.*, vol. 6, no. May 2014, p. 6370, 2015.

[92] C. Y. McLean *et al.*, “GREAT improves functional interpretation of cis-regulatory regions.,” *Nat. Biotechnol.*, vol. 28, no. 5, pp. 495–501, May 2010.

[93] R. C. Gentleman *et al.*, “Bioconductor: open software development for computational biology and bioinformatics.,” *Genome Biol.*, vol. 5, no. 10, p. R80, Jan. 2004.

[94] W. Huber *et al.*, “Orchestrating high-throughput genomic analysis with Bioconductor,” *Nat. Methods*, vol. 12, no. 2, pp. 115–121, Jan. 2015.

[95] B. Langmead and S. L. Salzberg, “Fast gapped-read alignment with Bowtie 2.,” *Nat. Methods*, vol. 9, no. 4, pp. 357–9, Apr. 2012.

[96] S. Durinck *et al.*, “BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis.,” *Bioinformatics*, vol. 21, no. 16, pp. 3439–40, Aug. 2005.

[97] S. Durinck, P. T. Spellman, E. Birney, and W. Huber, “Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt.,” *Nat. Protoc.*, vol. 4, no. 8, pp. 1184–91, Jan. 2009.

[98] S. Feng *et al.*, “Conservation and divergence of methylation patterning in plants and animals,” *Proc. Natl. Acad. Sci.*, vol. 107, no. 19, pp. 8689–8694, May 2010.

[99] A. Zemach, I. E. McDaniel, P. Silva, and D. Zilberman, “Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation,” *Science (80-.).*, vol. 328, no. 5980, pp. 916–919, May 2010.

[100] C. E. Niederhuth *et al.*, “Widespread natural variation of DNA methylation within angiosperms,” *Genome Biol.*, vol. 17, no. 194, 2016.

[101] S. Takuno, J.-H. Ran, and B. S. Gaut, “Evolutionary patterns of genic DNA methylation vary across land plants,” *Nat. Plants*, vol. 2, no. January, p. 15222, 2016.

[102] J. A. Law and S. E. Jacobsen, “Establising, maintaining and modifying DNA methylation patterns in plants and animals,” *Nat Rev Genet.*, vol. 11, no. 3, pp. 204–220, 2010.

[103] M. a Matzke, T. Kanno, and A. J. M. Matzke, “RNA-Directed DNA Methylation: The Evolution of a Complex Epigenetic Pathway in Flowering Plants.,” *Annu. Rev. Plant Biol.*, no. December 2014, pp. 1–25, 2014.

[104] S. Cortijo *et al.*, “Mapping the Epigenetic Basis of Complex Traits,” *Science (80-.).*, vol. 343, no. 6175, pp. 1145–1148, Mar. 2014.

[105] F. Johannes *et al.*, “Assessing the impact of transgenerational epigenetic variation on complex traits,” *PLoS Genet.*, vol. 5, no. 6, 2009.

[106] J. Reinders *et al.*, “Compromised stability of DNA methylation and transposon immobilization in mosaic Arabidopsis epigenomes,” *Genes Dev.*, vol. 23, no. 8, pp. 939–950, 2009.

[107] M. Mirouze *et al.*, “Loss of DNA methylation affects the recombination landscape in Arabidopsis.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 15, pp. 5880–5885, 2012.

[108] N. E. Yelina, C. Lambing, T. J. Hardcastle, X. Zhao, B. Santos, and I. R. Henderson, “DNA methylation epigenetically silences crossover hot spots and controls chromosomal domains of meiotic recombination in Arabidopsis.,” *Genes Dev.*, vol. 29, no. 20, pp. 2183–202, Oct. 2015.

[109] M. Colome-Tatche *et al.*, “Features of the Arabidopsis recombination landscape resulting from the combined loss of sequence variation and DNA methylation,” *Proc. Natl. Acad. Sci.*, vol. 109, no. 40, pp. 16240–16245, 2012.

[110] C. Melamed-Bessudo and a. a. Levy, “PNAS Plus: Deficiency in DNA methylation increases meiotic crossover rates in euchromatic but not in heterochromatic regions in Arabidopsis,” *Proc. Natl. Acad. Sci.*, vol. 109, no. 16, pp. E981--E988, 2012.

[111] S. Tsukahara, A. Kobayashi, A. Kawabe, O. Mathieu, A. Miura, and T. Kakutani, “Bursts of retrotransposition reproduced in Arabidopsis.,” *Nature*, vol. 461, no. 7262, pp. 423–426, 2009.

[112] M. Mirouze *et al.*, “Selective epigenetic control of retrotransposition in Arabidopsis.,” *Nature*, vol. 461, no. September, pp. 1–5, 2009.

[113] A. Miura, S. Yonebayashi, K. Watanabe, T. Toyama, H. Shimada, and T. Kakutani, “Mobilization of transposons by a mutation abolishing full DNA methylation in Arabidopsis,” *Nature*, vol. 411, no. May, pp. 212–214, 2001.

[114] T. Singer, C. Yordan, and R. A. Martienssen, “Robertson’s Mutator transposons in A. thaliana are regulated by the chromatin-remodeling gene Decrease in DNA Methylation (DDM1),” *Genes Dev.*, vol. 15, no. 5, pp. 591–602, 2001.

[115] C. Cheng *et al.*, “Loss of function mutations in the rice chromomethylase OsCMT3a cause a burst of transposition,” *Plant J.*, vol. 83, no. 6, pp. 1069–1081, 2015.

[116] D. Secco *et al.*, “Stress induced gene expression drives transient DNA methylation changes at adjacent repetitive elements.,” *Elife*, vol. 4, no. July, p. e09343, 2015.

[117] X. Zhang, “Dynamic differential methylation facilitates pathogen stress response in Arabidopsis,” *Proc. Natl. Acad. Sci.*, vol. 109, no. 32, pp. 12842–12843, 2012.

[118] A. Yu *et al.*, “Dynamics and biological relevance of DNA demethylation in Arabidopsis antibacterial defense.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 6, pp. 2389–2394, 2013.

[119] A. López Sánchez, J. H. M. Stassen, L. Furci, L. M. Smith, and J. Ton, “The role of DNA (de)methylation in immune responsiveness of Arabidopsis,” *Plant J.*, vol. 88, no. 3, pp. 361–374, Nov. 2016.

[120] R. J. Schmitz *et al.*, “Transgenerational Epigenetic Instability Is a Source of Novel Methylation Variants,” *Science (80-.).*, vol. 334, no. 6054, pp. 369–373, Oct. 2011.

[121] C. Becker *et al.*, “Spontaneous epigenetic variation in the Arabidopsis thaliana methylome,” *Nature*, vol. 480, no. 7376, pp. 245–249, Sep. 2011.

[122] C. Jiang, A. Mithani, E. J. Belfield, R. Mott, L. D. Hurst, and N. P. Harberd, “Environmentally responsive genome-wide accumulation of de novo Arabidopsis thaliana mutations and epimutations.,” *Genome Res.*, vol. 24, no. 11, pp. 1821–9, Nov. 2014.

[123] L. Quadrana and V. Colot, “Plant Transgenerational Epigenetics,” *Annu. Rev. Genet.*, vol. 50, no. 1, pp. 467–491, Nov. 2016.

[124] A. Vidalis, D. Živković, R. Wardenaar, D. Roquis, A. Tellier, and F. Johannes, “Methylome evolution in plants,” *Genome Biol.*, vol. 17, no. 1, p. 264, 2016.

[125] C. M. Diez, K. Roessler, and B. S. Gaut, “Epigenetics and plant genome evolution,” *Curr. Opin. Plant Biol.*, vol. 18, no. 1, pp. 1–8, 2014.

[126] D. Weigel and V. Colot, “Epialleles in plant evolution,” *Genome Biol.*, vol. 13, no. 10, p. 249, 2012.

[127] N. M. Springer, “Epigenetics and crop improvement,” *Trends Genet.*, vol. 29, no. 4, pp. 241–247, 2013.

[128] L. Ji, D. A. Neumann, and R. J. Schmitz, “Crop Epigenomics: Identifying, Unlocking, and Harnessing Cryptic Variation in Crop Genomes,” *Mol. Plant*, vol. 8, no. 6, pp. 860–870, 2015.

[129] E. J. Finnegan, W. J. Peacock, and E. S. Dennis, “Reduced DNA methylation in Arabidopsis thaliana results in abnormal plant development.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 93, no. 16, pp. 8449–54, Aug. 1996.

[130] A. M. Lindroth *et al.*, “Requirement of CHROMOMETHYLASE3 for Maintenance of CpXpG Methylation,” *Science (80-.).*, vol. 292, no. 5524, 2001.

[131] J. Du *et al.*, “Dual Binding of Chromomethylase Domains to H3K9me2-Containing Nucleosomes Directs DNA Methylation in Plants,” *Cell*, vol. 151, no. 1, pp. 167–180, Sep. 2012.

[132] H. Stroud *et al.*, “Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis,” *Nat. Struct. Mol. Biol.*, vol. 21, no. 1, pp. 64–72, Dec. 2013.

[133] X. Cao and S. E. Jacobsen, “Locus-specific control of asymmetric and CpNpG methylation by the DRM and CMT3 methyltransferase genes,” *Proc. Natl. Acad. Sci.*, vol. 99, no. Supplement 4, pp. 16491–16498, Dec. 2002.

[134] X. Cao and S. E. Jacobsen, “Role of the arabidopsis DRM methyltransferases in de novo DNA methylation and gene silencing.,” *Curr. Biol.*, vol. 12, no. 13, pp. 1138–44, Jul. 2002.

[135] C. Alonso, R. PÃ©rez, P. Bazaga, and C. M. Herrera, “Global DNA cytosine methylation as an evolving trait: phylogenetic signal and correlated evolution with genome size in angiosperms,” *Front. Genet.*, vol. 6, p. 4, Jan. 2015.

[136] T. Kawakatsu *et al.*, “Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions,” *Cell*, vol. 166, no. 2, pp. 492–506, 2016.

[137] H. Stroud, M. V. C. Greenberg, S. Feng, Y. V Bernatavichute, and S. E. Jacobsen, “Comprehensive Analysis of Silencing Mutants Reveals Complex Regulation of the Arabidopsis Methylome,” *Cell*, vol. 152, no. 17, pp. 352–364, 2013.

[138] A. Taudt, D. Roquis, A. Vidalis, R. Wardenaar, F. Johannes, and M. Colome-Tatché, “METHimpute: imputation-guided construction of complete methylomes from WGBS data,” *BMC Genomics*, vol. 19, no. 1, p. 444, Dec. 2018.

[139] F. Krueger and S. R. Andrews, “Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications,” *Bioinformatics*, vol. 27, no. 11, pp. 1571–1572, 2011.

[140] A. Akalin *et al.*, “methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles.,” *Genome Biol.*, vol. 13, no. 10, p. R87, 2012.

[141] W. Guo *et al.*, “BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data,” *BMC Genomics*, vol. 14, no. 1, p. 774, 2013.

[142] Q. Gouil and D. C. Baulcombe, “DNA Methylation Signatures of the Plant Chromomethyltransferases,” *PLOS Genet.*, vol. 12, no. 12, p. e1006526, Dec. 2016.

[143] M. Martin, “Cutadapt removes adapter sequences from high-throughput sequencing reads,” *EMBnet.journal*, vol. 17, no. 1, pp. 10–12, 2011.

[144] B. Langmead and S. L. Salzberg, “Fast gapped-read alignment with Bowtie 2,” *Nat Methods*, vol. 9, no. 4, pp. 357–359, 2012.

[145] H. Li *et al.*, “The Sequence Alignment/Map format and SAMtools,” *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[146] H. Stroud *et al.*, “Plants regenerated from tissue culture contain stable epigenome changes in rice,” *Elife*, vol. 2013, no. 2, pp. 1–14, 2013.

[147] M. Regulski *et al.*, “The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA,” *Genome Res.*, vol. 23, no. 10, pp. 1651–1662, 2013.

[148] The Arabidopsis Genome Initiative, “Analysis of the genome sequence of the flowering plant Arabidopsisthaliana,” *Nature*, vol. 408, no. 6814, pp. 796–815, Dec. 2000.

[149] I. R. G. Sequencing Project, “The map-based sequence of the rice genome,” *Nature*, vol. 436, no. 7052, pp. 793–800, Aug. 2005.

[150] T. Rice Annotation Project *et al.*, “Curated genome annotation of Oryza sativa ssp. japonica and comparative genome analysis with Arabidopsis thaliana.,” *Genome Res.*, vol. 17, no. 2, pp. 175–83, Feb. 2007.

[151] P. S. Schnable *et al.*, “The B73 Maize Genome: Complexity, Diversity, and Dynamics,” *Science (80-.).*, vol. 326, no. 5956, pp. 1112–1115, Nov. 2009.

[152] P. T. West *et al.*, “Genomic distribution of H3K9me2 and DNA methylation in a maize genome,” *PLoS One*, vol. 9, no. 8, pp. 1–10, 2014.

[153] R. Lister *et al.*, “Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis,” *Cell*, vol. 133, no. 3, pp. 523–536, 2008.

[154] S. J. Cokus *et al.*, “Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning,” *Nature*, vol. 452, no. 7184, pp. 215–219, 2008.

[155] E. Libertini *et al.*, “Saturation analysis for wholegenome bisulfite sequencing data,” *Nat. Publ. Gr.*, pp. 11–13, 2016.

[156] E. E. Holmes *et al.*, “Performance Evaluation of Kits for Bisulfite-Conversion of DNA from Tissues, Cell Lines, FFPE Tissues, Aspirates, Lavages, Effusions, Plasma, Serum, and Urine,” *PLoS One*, vol. 9, no. 4, p. e93933, Apr. 2014.

[157] C. A. Leontiou, M. D. Hadjidaniel, P. Mina, P. Antoniou, M. Ioannides, and P. C. Patsalis, “Bisulfite Conversion of DNA: Performance Comparison of Different Kits and Methylation Quantitation of Epigenetic Biomarkers that Have the Potential to Be Used in Non-Invasive Prenatal Testing.,” *PLoS One*, vol. 10, no. 8, p. e0135058, 2015.

[158] D. P. Genereux, W. C. Johnson, A. F. Burden, R. Stoger, and C. D. Laird, “Errors in the bisulfite conversion of DNA: modulating inappropriate- and failed-conversion frequencies,” *Nucleic Acids Res.*, vol. 36, no. 22, pp. e150–e150, Dec. 2008.

Appendix

List of abbreviations

5mC	cytosine methylation (5-methylcytosine)
aCGH	array comparative genomic hybridization
AD	Alzheimer's disease
AUC	area under curve
CDF	cumulative density function
ChIP-seq	chromatin immunoprecipitation followed by sequencing
CIN	chromosomal instability
CNA	copy number alteration
CNB	copy number breakpoint
CNV	copy number variation
eCDF	empirical cumulative density function
GEO	Gene expression omnibus
GML	genomewide methylation level
HMM	Hidden Markov Model
ITH	intra-tumor heterogeneity
KDE	kernel density estimation
mSFS	methylation site frequency spectrum
NGS	next generation sequencing
nls	non-linear least squares
PCR	polymerase chain reaction
QTL	quantitative trait locus
ROC	receiver operator characteristic
SAC	spindle assembly checkpoint
SCE	sister chromatid exchange
SCLC	small cell lung cancer
scWGS	single-cell whole genome sequencing
SoS	sum-of-squares
Strand-seq	single-cell DNA template strand sequencing
TE	transposable element
TSS	transcription start site
UCSC	University of California, Santa Cruz
WGBS	whole genome bisulfite sequencing
wGML	weighted genomewide methylation level

List of publications

Taudt, A., Roquis, D., Vidalis, A., Wardenaar, R., Johannes, F. & Colome-Tatché, M. **METHimpute: imputation-guided construction of complete methylomes from WGBS data.** *BMC Genomics* 19, 444 (2018).

Roquis, D., Taudt, A., Geyer, K. K., Padalino, G., Hoffmann, K. F., Holroyd, N., Berriman, M., Aliaga, B., Chaparro, C., Grunau, C. & Augusto, R. de C. **Histone methylation changes are required for life cycle progression in the human parasite Schistosoma mansoni.** *PLOS Pathog.* 14, e1007066 (2018).

Hanna, C. W., Taudt, A., Huang, J., Gahurova, L., Kranz, A., Andrews, S., Dean, W., Stewart, A. F., Colomé-Tatché, M. & Kelsey, G. **MLL2 conveys transcription-independent H3K4 trimethylation in oocytes.** *Nat. Struct. Mol. Biol.* 1 (2018).

Ferronika, P., van den Bos, H., Taudt, A., Spierings, D. C. J., Saber, A., Hiltermann, T. J. N., Kok, K., Porubsky, D., van der Wekken, A. J., Timens, W., Foijer, F., Colomé-Tatché, M., Groen, H. J. M., Lansdorp, P. M. & van den Berg, A. **Copy number alterations assessed at the single-cell level revealed mono- and polyclonal seeding patterns of distant metastasis in a small cell lung cancer patient.** *Ann. Oncol.* (2017).

Kebede, A. F., Nieborak, A., Shahidian, L. Z., Le Gras, S., Richter, F., Gómez, D. A., Baltissen, M. P., Meszaros, G., Magliarelli, H. de F., Taudt, A., Margueron, R., Colomé-Tatché, M., Ricci, R., Daujat, S., Vermeulen, M., Mittler, G. & Schneider, R. **Histone propionylation is a mark of active chromatin.** *Nat. Struct. Mol. Biol.* nsmb.3490 (2017).

Taudt, A., Colomé-Tatché, M. & Johannes, F. **Genetic sources of population epigenomic variation.** *Nat. Rev. Genet.* 17, 319–332 (2016).

van den Bos, H., Spierings, D. C. J., Taudt, A. S., Bakker, B., Porubský, D., Falconer, E., Novoa, C., Halsema, N., Kazemier, H. G., Hoekstra-Wakker, K., Guryev, V., den Dunnen, W. F. A., Foijer, F., Tatché, M. C., Boddeke, H. W. G. M. & Lansdorp, P. M. **Single-cell whole genome sequencing reveals no evidence for common aneuploidy in normal and Alzheimer's disease neurons.** *Genome Biol.* 17, 116 (2016).

Bakker, B., Taudt, A., Belderbos, M. E., Porubsky, D., Spierings, D. C. J., de Jong, T. V., Halsema, N., Kazemier, H. G., Hoekstra-Wakker, K., Bradley, A., de Bont, E. S. J. M., van den Berg, A., Guryev, V., Lansdorp, P. M., Colomé-Tatché, M. & Foijer, F. **Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies.** *Genome Biol.* 17, 115 (2016).

Taudt, A., Nguyen, M. A., Heinig, M., Johannes, F. & Colome-Tatche, M. **chromstaR: Tracking combinatorial chromatin state dynamics in space and time.** *bioRxiv* (Cold Spring Harbor Labs Journals, 2016).

van der Graaf, A., Wardenaar, R., Neumann, D. A., Taudt, A., Shaw, R. G., Jansen, R. C., Schmitz, R. J., Colomé-Tatché, M. & Johannes, F. **Rate, spectrum, and evolutionary dynamics of spontaneous epimutations.** *Proc. Natl. Acad. Sci. U. S. A.* 112, 6676–81 (2015).

Heinig, M., Colomé-Tatché, M., Taudt, A., Rintisch, C., Schafer, S., Pravenec, M., Hubner, N., Vingron, M. & Johannes, F. **histoneHMM: Differential analysis of histone modifications with broad genomic footprints.** *BMC Bioinformatics* 16, 60 (2015).

Taudt, A., Arnold, A. & Pleiss, J. **Simulation of protein association: Kinetic pathways towards crystal contacts.** *Phys. Rev. E* 91, (2015).

Kratzer, K., Berryman, J. T., Taudt, A., Zeman, J. & Arnold, A. **The Flexible Rare Event Sampling Harness System (FRESHS).** *Comput. Phys. Commun.* 185, 1875–1885 (2014).

Summary for laymen

Modern day biology increasingly relies on the production and analysis of huge amounts of digital data. This data is produced by new experimental techniques which allow ever deeper insights into the mechanisms by which our cells function. One such technique is “Next Generation Sequencing”. It was developed in the 2000s and is now widely applied to the study of all phenomena which involve the DNA sequence. It is also called a “molecular microscope”, because it enables researchers to study the sequence of DNA with extremely high resolution, and as easily as with a normal light microscope. There is, however, an important difference: While a normal microscope produces images which can be interpreted by the human eye, a sequencing machine produces sequences of letters. These letters are A, T, C and G, which are the basic building blocks of DNA. Our human genome contains approximately 3 billion of these letters, which equals approximately 130 printed books. This is the reason why we need computers to make sense out of this data.

The presented thesis describes novel algorithms for the analysis of Next Generation Sequencing (NGS) data. More specifically, four different algorithms are presented. These algorithms were developed for different varieties of NGS experiments, each of which allows a unique perspective into the cell.

Chapter 1 gives an introduction to Next Generation Sequencing and explains in more detail how the same technology can be used in a modified form to investigate different types of sequence related phenomena. Chapter 1 also explains the idea behind the algorithmic model that is used to analyze NGS data in this thesis, called Hidden Markov Model.

Chapter 2 presents a model for the analysis of copy number gains and losses in single cells from NGS data. A normal human cell has two copies of each chromosome, one originating from the father and one from the mother. Aberrations from this healthy physiological state of two copies cause neurological and developmental defects. The most prominent example for such a whole-chromosome copy number aberration, also called aneuploidy, is Down’s syndrome, where chromosome number 21 is present in three copies instead of two. Most other aneuploidies (whole-chromosome copy number aberration) are lethal. Aneuploidies are also very common in cancer cells, but in this case they are not lethal for the cancer cell but give them a growth advantage. Single-cell whole genome sequencing is an NGS-based technique which allows to study these copy number aberrations in individual cells. I present a Hidden Markov Model which can predict the copy number state from this data, and this method has been applied to investigate the role of aneuploidy in Alzheimer’s disease, as well as the role of smaller copy number aberrations in cancer cells. In Chapter 3 an extension of the model from Chapter 2 is presented. This extension allows the detection of copy number aberrations with much better resolution, and it is applicable to another NGS technique, called Strand-seq. Strand-seq also allows the study of the inheritance patterns of DNA during cell division.

Chapter 4 and 5 present models for the analysis of epigenetic modifications on the DNA. The term “epi” in epigenetics comes from Greek and means “on top of” or “in addition to”. Epigenetics thus refers to information “in addition to” the DNA sequence and molecular factors “on top of” the DNA sequence. All cells in an organism have the identical DNA sequence, but still, a heart cell is quite different from a brain cell or a skin cell. Epigenetic factors are responsible for these differences, such as chemical modifications on the proteins around which the DNA is packed (histone modifications) or chemical modifications of the DNA itself (DNA methylation). Again, Next Generation Sequencing can be used to map these epigenetic factors, and I present Hidden Markov Models for the analysis of histone modifications and DNA methylation. These models can help to increase our knowledge of how totipotent stem cells differentiate into specialized cell types, and how the environment can affect gene expression.

Summary for laymen (in Dutch)

Onderzoekers in de hedendaagse biologie zijn steeds meer afhankelijk van de productie en analyse van grote hoeveelheden digitale data. Deze data wordt geproduceerd met behulp van nieuwe experimentele technieken die grotere inzichten verschaffen over de manier waarop onze cellen werken. Eén van deze technieken is next-generation sequencing (NGS). Deze techniek werd ontwikkeld in het begin van de 21e eeuw, en wordt nu op brede wijze toegepast om alles omtrent de basepaarvolgorde (sequentie) van het DNA te bestuderen. NGS wordt ook wel "moleculaire microscopie" genoemd omdat het onderzoekers in staat stelt om met hoge resolutie de DNA-sequentie te bestuderen, en wel met het gemak van een gewone lichtmicroscoop. Er is echter een belangrijk verschil: terwijl een lichtmicroscoop beelden genereert die kunnen worden geïnterpreteerd door het menselijk oog, genereert een NGS-machine een reeks letters (wat ook wel sequencing wordt genoemd). Deze letters zijn A, T, C en G, welke de bouwstenen vormen van het DNA. Het menselijke genoom bevat circa 3 miljard van deze letters. Wanneer deze letters zouden worden opgeschreven kunnen er ongeveer 130 boeken gevuld worden. Om deze reden heeft men computers nodig om dit soort data te begrijpen.

Dit proefschrift beschrijft vier nieuwe algoritmes voor de analyse van NGS-data, welke zijn ontwikkeld voor verschillende NGS-experimenten die elk een uniek inzicht leveren in de cel.

Hoofdstuk 1 dient als introductie voor NGS en legt in detail uit hoe dezelfde technologie gebruikt kan worden om verschillende soorten fenomenen gerelateerd aan de DNA-sequentie te onderzoeken. Hoofdstuk 1 weidt ook uit over het concept achter het algoritmisch model dat is gebruikt om NGS-data te analyseren, het zogeheten Hidden Markov Model.

Hoofdstuk 2 behandelt een model voor de analyse van kopienummerveranderingen in individuele cellen aan de hand van NGS-data. Een normale menselijke cel bevat twee kopieën van elk chromosoom, één kopie van de vader en één van de moeder. Veranderingen van deze fysiologische staat van twee kopieën veroorzaakt neurologische en ontwikkelingsstoornissen. Het meest bekende voorbeeld van zo een afwijking in het chromosoomaantal, ook wel aneuploïdie genoemd, is het syndroom van Down, waarbij drie kopieën van chromosoom 21 aanwezig zijn. De meeste andere vormen van aneuploïdie zijn dodelijk. Aneuploïdie komt zeer vaak voor in kankercellen, maar is in deze cellen niet dodelijk en geeft veelal een groeivoordeel. Het sequencen van individuele cellen met behulp van NGS-technologie (single-cell whole genome sequencing) stelt men in staat om het aantal chromosoomkopieën te bepalen in afzonderlijke cellen. Hier beschrijf ik een Hidden Markov Model dat deze kopienummers kan voorspellen aan de hand van NGS-data, en deze methode is toegepast om de rol van aneuploïdie in de ziekte van Alzheimer te bestuderen en de rol van kleine kopienummerveranderingen in kankercellen. In hoofdstuk 3 wordt uitgeweid over een uitbreiding van het in hoofdstuk 2 beschreven model. Deze uitbreiding kan kopienummerveranderingen met een hogere resolutie detecteren en is toepasbaar op een andere NGS-techniek, het zogeheten 'Strand-seq'. Strand-seq kan worden gebruikt om de overervingspatronen van het DNA tijdens de celdeling te bestuderen.

Hoofdstukken 4 en 5 beschrijven modellen voor de analyse van epigenetische wijzigingen van het DNA. Het voorvoegsel ‘epi’ in epigenetica is afkomstig uit het Grieks en betekent ‘bovenop’ of ‘naast’. Epigenetica verwijst dus naar informatie naast en moleculaire factoren bovenop de DNA-sequentie. Elke cel in een organisme heeft dezelfde DNA-sequentie, en toch is een hartcel zeer verschillend van een hersencel of een huidcel. Epigenetische factoren zijn verantwoordelijk voor deze verschillen, zoals chemische wijzigingen van de eiwitten waaromheen het DNA is gewikkeld (histonwijzigingen) of wijzigingen van het DNA zelf (DNA-methylering). NGS kan gebruikt worden om ook deze epigenetische factoren in kaart te brengen, en ik beschrijf hier een Hidden Markov Model voor de analyse van histonwijzigingen en DNA-methylering. Deze modellen kunnen helpen om onze kennis te vergroten over hoe totipotente stamcellen differentiëren in gespecialiseerde celtypen, en hoe de omgeving de expressie van genen kan beïnvloeden.

Acknowledgements

The past four years of my PhD were the background for tremendous personal and professional change in my life. I have met many people during that time, often inspiring, sometimes challenging, but all amazing humans. I would like to thank the most important ones here.

First of all, I want to thank **Maria Colomé-Tatché** for being the best supervisor a PhD student can wish for. I always felt that my scientific growth was your core interest, and the actual work that I produced was only secondary. I thank you for your perseverance when I wanted to give up. I appreciate the support for conferences that were outside of my PhD work. I value that you always had time to discuss with me and listened to my opinions.

I want to express my gratitude to other people that were supportive of my work. I thank my co-supervisor **Frank Johannes** for good discussions and the opportunity for a nice review. I thank **Peter Lansdorp** for his encouragement to develop software for Strand-seq data and a very good bottle of wine.

This thesis is a lot to read and to understand, and I want to thank the members of my assessment committee, **Jan Korbel**, **Bart Eggen** and **Stephan Ossowski**, for their time and interest in assessing this thesis. I hope they found this lecture interesting and entertaining.

I am grateful that I could collaborate with outstanding scientists in my time at ERIBA. Thank you **Bjorn**, **Diana**, **David** and **Hilda**. You all had a huge share in the success of my thesis and working with you was really fun. I appreciate your awesome experimental work and your enthusiasm in filing bug reports for AneuFinder. I also want to thank my collaborator **Courtney** at the Babrahm Institute for making me feel welcome there for one month and an outstanding collaboration.

What made my time in Groningen really enjoyable were my friends there and the activities we did together. Thank you **Clémence** for being my first friend in Groningen, our many walks and the wonderful food at your place. **Céline**, **Stijn**, **Sonia** and **Clémence**, I very much enjoyed our common Tango lessons and often felt much better afterwards than before.

I am going to miss my veteran WAMPEX buddies **Bjorn** and **Tristan** and all the others and the many hours we spent wandering around in fields at night, freezing, tired and feeling lost. Again any time ;) Also thank you **Bjorn** for joining me for the Oktoberfest and hosting me for my visit in Groningen, it was really fun. And a special thanks for translating my summary to Dutch. **Daniele**, I am glad we found so many opportunities to run together, play Squash, or simply eat and talk.

Also my time in Munich was full of fantastic people. **Dr. Oquis** and **Mamaryllis**, it was a pleasure to share an office with you. You are the coolest office mates ever! Thank you for all the board game evenings, drinks and support when I needed it the most. **Akshaya**, I was happy to have another PhD student in my group when you joined, and it was very nice to share a flat with you.

I also want to thank **Akshaya**, **Johanna** and **Anna** for being great colleagues and creating an enjoyable atmosphere at the ICB office. Also thanks to the other people at ICB for countless birthday parties, Beach Volleyball and Badminton sessions. Thank you **Lisa**, **Valle**, **Atefeh**, **David** and all the others.

Alejandra, thank you for asking me to run a Marathon with you. That was one of coolest things I ever did. And for having a Bachata lesson with me. But most of all, thank you for everything else that you have done for me. THANK YOU

A big, big THANK YOU goes to my paranymphs **Daniele** and **David**. I am grateful that you agreed to this honorable task and I feel lucky to have such amazing friends supporting me during my defense.

Vielen Dank der Winnengang, **Yogi**, **Sky** und **Dani**. Unsere Urlaube während meines Doktors waren legendär, und unsere Freundschaft wird jedes Jahr tiefer. DANKE

Und last but not least, will ich mich bei meiner Familie bedanken für die Unterstützung und Besuche während meiner Zeit in Groningen und München. Danke **Simon** und **Debbie** für euren Besuch in Groningen und dafür dass ihr super coole Geschwister seid! Danke **Christian** für dein Auto, deine Hilfe beim Umzug nach und von Groningen, und für die Radtour letztes Jahr. Danke dass ich jederzeit bei dir vorbeikommen kann. Danke **Tina** für deine Hilfe beim Umziehen, für die Plätzchen zu Weihnachten und überhaupt für alles andere.